SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 1–20, January 1995

A PARABOLIC EQUATION OF THE KPP TYPE IN HIGHER DIMENSIONS*

JEAN-FRANÇOIS MALLORDY† AND JEAN-MICHEL ROQUEJOFFRE‡

Abstract. The aim of this paper is to study the long-time behaviour of the solutions of a semilinear parabolic equation in an infinite cylinder, which generalises a well-known one-dimensional model that was first investigated by Kolmogorov, Petrovskii, and Piskunov (KPP).

It is shown that the asymptotic behaviour of the solutions strongly depends on the asymptotic behaviour of the initial datum at the ends of the cylinder. In particular, the equation admits a continuum of travelling wave solutions, each of them stable with regard to adequate perturbations.

Key words. reaction diffusion, KPP nonlinearity, travelling waves, asymptotic behavior

AMS subject classifications. 35B35, 35K57, 35B40

1. Introduction.

This paper is concerned with the large-time behaviour of the solutions of a class of semilinear parabolic equations in an infinite multidimensional cylinder. Namely, denoting by Σ the cylinder $\Sigma = (x, y) \in \mathbb{R} \times \omega$, ω being a bounded regular domain in $\mathbb{R}^{N-1}$, we consider the following problem:

$$(\mathcal{P}) : \begin{cases} \partial_t u - \Delta u + \alpha(y)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 \quad \text{on } \partial\Sigma, \\ u(t=0) = u_0 & \text{in } \Sigma, \end{cases}$$

Here ν denotes the outward unit normal on $\partial\Sigma$, $\alpha(y)$ is a given continuous function on $\bar{\omega}$, and g is a given source term satisfying $g(0) = g(1) = 0$ with $g > 0$ in $(0,1)$, and $g'(0) > 0 > g'(1)$. The initial datum u_0 is assumed to take its values in $(0,1)$.

Under the above assumptions, $(\mathcal{P})$ has a unique global classical solution, denoted by $S(t)u_0$. These assumptions will always be understood to hold in the rest of the paper.

Our aim is to understand how the asymptotics of u, for large t, depend on the initial datum u_0. Indeed, $(\mathcal{P})$ admits travelling wave solutions of speed $c \geq c_*$, for some c_* depending on the source term g, and it is expected that each travelling wave is stable in some sense.

Such problems have a wide range of applications, such as population dynamics or combustion; quite a few authors have already contributed to the study of the one-dimensional problem

$$u_t - u_{xx} = g(u), \quad u(-\infty, t) = 0, \quad u(+\infty, t) = 1.$$

The first paper dealing with the subject is the celebrated article by Kolmogorov, Petrovskii, and Piskunov [13]. The source term which is taken into account in their paper is $g(u) = u(1-u)$ and $S(t)H$—H being the Heaviside step function—is shown to converge, in some weak sense, towards the travelling front with speed c_*.

Subsequent work was then done by Aronson and Weinberger [2], where, in particular, the set of c's such that $u(x - ct, t)$ goes to 0 (respectively, 1) as $t \to +\infty$ is

* Received by the editors March 26, 1993; accepted for publication (in revised form) July 29, 1993.

† 31, avenue de la Libération, Clermont–Ferrand, France.

‡ Centre National de la Recherche Scientifique, Ecole Polytechnique, Centre de Mathématiques Appliquées 91128, Palaiseau Cedex, France.

characterised. Local exponential stability results were proved by Sattinger [19] for the waves with speed $c > c_*$ in weighted spaces; quite recently, Kirchgässner [12] reached local algebraic stability for the waves with speed c_*. In [10], Hagan investigated fully nonlinear equations of the form $u_t = f(u, u_x, u_{xx})$, with f satisfying suitable assumptions.

Finally, let us mention the work of Uchiyama [21], which covers all the previously mentioned aspects, and which provides a classification of every possible asymptotic behaviour.

In the multidimensional framework, Aronson and Weinberger [3] investigated the unsteady problem in $\mathbb{R}^N$, with $\alpha \equiv 0$. They obtained "hair-trigger effect" type results, i.e., a nonnegative initial datum with even a small compact support that evolves into a solution that tends to 1 on every compact subset.

Our purpose is to extend to $(\mathcal{P})$ some of these results. In particular, we will fully generalise [2], [19], and the global stability results of [21] for the waves with speed $c > c_*$. To do this, we will rely heavily on the results of Berestycki and Nirenberg [8].

The paper is organised as follows. In §2, we recall the existence, uniqueness, and qualitative properties results for the travelling wave solutions $(\mathcal{P})$. In §§3 and 4, we investigate $(\mathcal{P})$ with a source term g such that $g(s)/s$ is decreasing in s. Such a g will be called a KPP source term, and we will see in §§3 and 4 that this additional assumption allows us to produce shorter and more elegant proofs. Section 5 is devoted to exponential local stability results, and no KPP assumption is made unless otherwise stated. Finally, §6 is devoted to the extension of the results of §4 when g is no longer a KPP source term.

2. Travelling waves. All the material in this section comes from the work of Berestycki and Nirenberg [8]. Therefore no proof will be given, apart from the last exponential behaviour result.

When we look for travelling waves solutions of the form $u(x, y, t) = \phi(x + ct, y)$, we have to deal with the elliptic problem

$$(E) : \begin{cases} -\Delta\phi + (\alpha(y) + c)\partial_x\phi = g(\phi) & \text{in } \Sigma, \\ \partial_\nu\phi = 0 & \text{on } \partial\Sigma. \end{cases}$$

We set

$$\beta(y) = \alpha(y) + c, \quad \forall y \in \bar{\omega},$$
$$A_c = -\Delta + \beta(y)\partial_x.$$

The resolution of problem (E) requires the study of the linearised problem around 0, which is presented below.

The linearised problem. Here we are interested in finding positive solutions to the problem

$$(E_L) : \begin{cases} -\Delta z + \beta(y)\partial_x z = g'(0)z & \text{in } \Sigma, \\ \partial_\nu z = 0 & \text{on } \partial\Sigma. \end{cases}$$

It is easy to see that $z = e^{\lambda x}\varphi(y)$, with $\varphi > 0$ on ω, is a solution of (E_L) if and only if

$$\lambda^2 = \mu_1(-\Delta + \lambda\beta(y) - g'(0)),$$

where $\mu_1(B)$ represents the first eigenvalue of the operator B defined on ω with a Neumann condition on the boundary.

LEMMA 2.1. *Let $m \in \mathbb{R}$ and $\mu_1(t) = \mu_1(-\Delta + t\beta(y) + m)$, then the function $t \mapsto \mu_1(t)$ satisfies the following:*

1. *μ_1 is Lipschitz with coefficient $||\beta||_\infty$,*
2. *$\mu_1(0) = m$,*
3. *μ_1 is concave,*
4. *μ_1 is differentiable and $\mu_1'(0) = \frac{1}{|\omega|}\int_\omega \beta(y)dy$,*
5. *$m + t(\inf_\omega \beta) \leq \mu_1(t) \leq m + t(\sup_\omega \beta)$.*

We call $\lambda \in \mathbb{R}$ a principal eigenvalue (p.e.) if $\mu_1(\lambda) = \lambda^2$. An immediate consequence of the lemma is the following theorem (remember $m = -g'(0) < 0$).

THEOREM 2.2. *There exists $c_0 \in \mathbb{R}$ such that*

- *$c < c_0 \Longrightarrow$ there is no positive p.e.*
- *$c = c_0 \Longrightarrow$ there is exactly one positive p.e.*
- *$c > c_0 \Longrightarrow$ there are exactly two positive p.e.*

Notation: we call λ_c the smallest positive principal eigenvalue associated to $c > c_0$ and λ_0 the positive principal eigenvalue associated to c_0. We will also use the notation λ_c^u (u for "usual") instead of λ_c, and then λ_c^a (a for "accidental") will be the other p.e.

Solutions of Problem (E).

THEOREM 2.3. *There exists $c^* \geq c_0$ such that the problem*

$$(E): \begin{cases} -\Delta\phi + (\alpha(y) + c)\partial_x\phi = g(\phi) & \text{in } \Sigma, \\ \partial_\nu\phi = 0 & \text{on } \partial\Sigma. \end{cases}$$

with $\phi(-\infty, y) = 0$, $\phi(+\infty, y) = 1$ uniformly on $\bar{\omega}$, has a solution if and only if $c \geq c^*$. Furthermore the solution is unique (up to the x translations), and increasing in x, more precisely, $\partial_x\phi > 0$ on $\bar{\Sigma}$.

Notation: we call ϕ_c the solution associated to c normalised by the condition $\sup_{y\in\bar{\omega}} \phi_c(0, y) = 1/2$. If we now impose $\phi(-\infty, y) = 1$, $\phi(+\infty, y) = 0$ uniformly on $\bar{\omega}$, we will have solutions if and only if $c \leq c^{**}$, for a real $c^{**} < c^*$, this is obvious by changing x into $-x$. We call $\tilde{\phi}_c$ these decreasing solutions. We call λ_{00} the λ_0 obtained when considering $-\alpha(y)$ instead of $\alpha(y)$ in the linearised problem.

We end this paragraph by examining the asymptotic behaviour of ϕ_c as $|x| \to +\infty$.

LEMMA 2.4.

- *If $c > c^*$ we have, near $-\infty$, and uniformly in $y \in \bar{\omega}$,*

$$\phi_c(x, y) \sim e^{\lambda_- x}\varphi_-(y), \qquad \partial_x\phi_c(x, y) \sim \lambda_- e^{\lambda_- x}\varphi_-(y);$$

  furthermore, if $c^* > c_0$, we have the same at $c = c^*$.
- *If $c \geq c^*$, we have near $+\infty$, and uniformly in $y \in \bar{\omega}$,*

$$1 - \phi_c(x, y) \sim e^{-\lambda_+ x}\varphi_+(y), \qquad \partial_x\phi_c(x, y) \sim \lambda_+ e^{-\lambda_+ x}\varphi_+(y),$$

where $e^{\lambda_- x}\varphi_-(y)$ is a positive solution of the linearised problem around 0 (E_L) with $\lambda_- > 0$, and $e^{-\lambda_+ x}\varphi_+(y)$ is a positive solution of the linearised problem around 1, with $\lambda_+ > 0$.

In the case $c > c^*$, we can be more precise.

LEMMA 2.5. *Let $c > c^*$, then the exponential behaviour of ϕ_c near $-\infty$ is given by λ_c, the smallest principal eigenvalue, i.e.,*

$$\phi_c(x, y) \sim e^{\lambda_c x}\varphi_c(y) \quad \text{as } x \to -\infty,$$

where φ_c is a positive eigenfunction associated to λ_c.

Proof. By the sliding method, as in [7]. Set $\phi_c(x,y) \sim \varphi_-(y)e^{\lambda_- x}$ as $x \to -\infty$, and assume that $\lambda_- > \lambda_c^a$. Then we have

$$\lim_{|x|\to+\infty} \phi_{c^*}/\phi_c = +\infty$$

(this follows from Theorem 4.1 in [8]). Moreover,

$$-\Delta(\phi_{c^*} - \phi_c) + (c + \alpha(y))\partial_x(\phi_{c^*} - \phi_c) + \frac{g(\phi_{c^*}) - g(\phi_c)}{\phi_{c^*} - \phi_c}(\phi_{c^*} - \phi_c)$$
$$= (c - c^*)\partial_x\phi_{c^*} > 0.$$

We choose a translate of ϕ_{c^*} (still denoted by ϕ_{c^*}), such that $\phi_{c^*} \geq \phi_c$, and such that the two functions coincide somewhere. The maximum principle and the Hopf boundary lemma yield $\phi_{c^*} = \phi_c$, which is a contradiction. □

3. Long-time behaviour with rapidly decreasing initial data. Here we consider the problem

$$(\mathcal{P}) : \begin{cases} \partial_t u - \Delta u + \alpha(y)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 & \text{on } \partial\Sigma, \\ u(t=0) = u_0 & \text{in } \Sigma, \end{cases}$$

where u_0 decays faster than any travelling front at one or both ends of the cylinder. We recall that, in this section, g is assumed to be a KPP source term.

3.1. Decay at only one end.

THEOREM 3.1.

- *If* $u_0(x,y) = o(e^{\lambda x})$ *as* $x \to -\infty$ *for every* $\lambda < \lambda^*$, *then for each* $c > c^*$ *we have* $u(x - ct, y, t) \to 0$ *when* $t \to +\infty$ *uniformly on each* $(-\infty, a] \times \bar{\omega}$.
- *If* $\liminf_{x\to+\infty} u_0 > 0$, *then for each* $c < c^*$ *we have* $u(x - ct, y, t) \to 1$ *when* $t \to +\infty$ *uniformly on each* $[a, +\infty) \times \bar{\omega}$.

If we now consider a u_0 that satisfies both conditions of Theorem 3.1, we can obviously define

$$\varphi(t) = \sup\{x \in \mathbb{R}; u(x,y,t) \leq 1/2, \forall y \in \omega\}$$

for sufficiently large values of t. We then have the next corollary.

COROLLARY 3.2. *Under the conditions of Theorem* 3.1, $\varphi(t)/t \to -c^*$ *as* $t \to +\infty$.

We first prove the first part of the theorem, which is much easier. Set $v(x,y,t) = u(x - ct, y, t)$ so that $\partial_t v + A_c v = g(v)$; assume $c > c^*$. Let $c' \in (c^*, c)$ and $e^{\lambda x}\varphi(y)$ a positive solution of the linearised problem $\partial_t v + A_{c'} v = g'(0)v$, with the Neumann boundary condition (such a solution exists since $c > c^*$). Using the condition at $-\infty$, we know $\exists x_0$ such that

$$x < x_0 \Longrightarrow u_0(x,y) < e^{\lambda x}\varphi(y)$$

so, for some large M, $u_0(x,y) < e^{\lambda(x+M)}\varphi(y)$ on Σ; now just apply the maximum principle to see that $v(x,y,t) < e^{\lambda(x+M-(c'-c^*)t)}\varphi(y)$ on Σ, $\forall t \geq 0$, wich yields the first assertion of Theorem 3.1.

Before proving the second part of the theorem, we need a lemma.

LEMMA 3.3. *Let* f *be a* C^1 *function such that* $f(0) = f(1) = 0$, $f'(0) > 0 > f'(1)$, *and* $f > 0$ *on* $(0,1)$; *we suppose that* u *satisfies* $\partial_t u + A_c u = f(u)$ *in* Σ *with*

Neumann boundary conditions, and $\liminf_{t\to+\infty} u(x,y,t) \geq \delta > 0$ *uniformly on each* $[a,+\infty)\times\bar{\omega}$, *then* $u(x,y,t) \to 1$ *when* $t \to +\infty$ *uniformly on each* $[a,+\infty)\times\bar{\omega}$.

Proof. There exists a sequence $t_n \to +\infty$ such that $\forall t \geq t_n$, $u(t) \geq h_n^0$ where

$$h_n^0(x,y) = \begin{cases} \delta/2 & \text{if } x > -n, \\ 0 & \text{if } x < -n, \end{cases}$$

we denote by $h_n(x,y,t)$ the solution of

$$\begin{cases} \partial_t h + A_c h = f(h) & \text{in } \Sigma, \\ \partial_\nu h = 0 & \text{on } \partial\Sigma, \\ h(t=0) = h_n^0 & \text{in } \Sigma. \end{cases}$$

By the maximum principle, we have $u(t+t_n) \geq h_n(t)$ $\forall t \geq 0$. As n increases, h_n^0 increases, and so $h_n(t)$ increases too (another consequence of the maximum principle). Since $h_n(t)$ is bounded by 1, we have $h_n(x,y,t) \to h_\infty(x,y,t)$ as $n \to +\infty$. Using classical parabolic estimates, we get $h_n \to h_\infty$, $\partial_t h_n \to \partial_t h_\infty$, $\partial_x h_n \to \partial_x h_\infty$, and $\Delta h_n \to \Delta h_\infty$, uniformly on each $[-a,+a]\times\bar{\omega}\times[1/T,T]$. This shows $\partial_t h_\infty + A_c h_\infty = f(h_\infty)$ with the Neumann boundary condition. Meanwhile $h_\infty(x,y,t=0) \geq \delta/2 > 0$. Obviously, $h_\infty(x,y,t) \to 1$ uniformly on Σ as $t \to +\infty$.

We conclude that $h_n(x,y,t) \to 1$ uniformly on each $[-a,+a]\times\bar{\omega}$ as n and $t \to +\infty$, but with $h_n(t)$ increasing in x (maximum principle), the convergence is in fact uniform on each $[-a,+\infty)\times\bar{\omega}$. By remembering $u(t+t_n) \geq h_n(t)$, $\forall t \geq 0$, we have our lemma. □

Now we prove the second assertion in Theorem 3.1; first, we suppose that g is concave. Set $g = 0$ outside of $[0,1]$. For $\epsilon > 0$, consider the following problem, with unknown (u,c):

$$(E_\epsilon): \begin{cases} -\Delta u + (\alpha(y)+c)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 & \text{on } \partial\Sigma, \\ u(-\infty,y) = -\epsilon \quad \text{and} \quad u(+\infty,y) = 1. \end{cases}$$

Setting $w = (u+\epsilon)/(1+\epsilon)$, we see that (E_ϵ) is equivalent to

$$(E'_\epsilon): \begin{cases} -\Delta w + (\alpha(y)+c)\partial_x w = g_\epsilon(w) & \text{in } \Sigma, \\ \partial_\nu w = 0 & \text{on } \partial\Sigma, \\ w(-\infty,y) = 0 \quad \text{and} \quad w(+\infty,y) = 1, \end{cases}$$

where we have set $g_\epsilon(s) = g((1+\epsilon)s-\epsilon)/(1+\epsilon)$. Using the concavity of g, we find that g_ϵ increases as ϵ decreases. Now, we just do as Berestycki and Nirenberg did when they proved that the existence of c^* (see [8]); this yields that for a given $\epsilon > 0$, there is exactly one solution (u_ϵ, c_ϵ) to problem (E_ϵ) (defined modulo x translations); this solution is increasing in x and c_ϵ increases to c^* as ϵ decreases to 0.

Let $\delta > 0$ such that $\liminf_{x\to+\infty} > \delta$; set $g_\delta(s) = \delta g(s/\delta)$ for $s \in \mathbb{R}$. Still using the concavity of g, we get $g_\delta \leq g$, $\forall \delta \in (0,1]$. Consider the problem

$$(E_{\epsilon,\delta}): \begin{cases} -\Delta v + (\alpha(y)+c)\partial_x v = g_\delta(v) & \text{in } \Sigma, \\ \partial_\nu v = 0 & \text{on } \partial\Sigma, \\ v(-\infty,y) = -\epsilon \quad \text{and} \quad v(+\infty,y) = \delta. \end{cases}$$

Setting $v = \delta u$, we see that problem $(E_{\epsilon,\delta})$ is equivalent to problem $(E_{\epsilon/\delta})$ (defined above); let us call $(v_{\epsilon,\delta}, c_{\epsilon,\delta})$ the solution of problem $(E_{\epsilon,\delta})$; we still have that $c_{\epsilon,\delta}$ increases to c^* as ϵ decreases to 0.

Let $c < c^*$. Let $\epsilon > 0$ such that $c < c_{\epsilon,\delta} < c^*$; we now consider u the solution of

$$(\mathcal{P}_C) : \begin{cases} \partial_t u - \Delta u + (\alpha(y) + c)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 \quad \text{on } \partial\Sigma, \\ u(t = 0) = u_0 & \text{in } \Sigma, \end{cases}$$

and we set $z(x, y, t) = v_{\epsilon,\delta}(x + (c_{\epsilon,\delta} - c)t, y)$. We have

$$\begin{cases} -\Delta z + (\alpha(y) + c)\partial_x z = g_\delta(z) \leq g(z) & \text{in } \Sigma, \\ \partial_\nu z = 0 \quad \text{on } \partial\Sigma, \\ z(x, y, t = 0) \leq u_0(x, y) \text{ in } \Sigma \text{ up to some } x \text{ translation}, \end{cases}$$

which implies $z(x, y, t) \leq u(x, y, t), \forall t \geq 0, \forall (x, y) \in \Sigma$. This yields $\liminf_{t\to+\infty} u(x, y, t) \geq \delta$ uniformly on each $[-a, +\infty) \times \bar{\omega}$. An application of Lemma 3.3 ends the proof, in the case where g is KPP and concave.

The result is still available when g is KPP but not concave: for $\eta > 0$, set $c^*(g'(0) - \eta)$ the critical speed obtained when replacing $g'(0)$ by $g'(0) - \eta$. Choose some small $\eta > 0$ such that $c < c^*(g'(0) - \eta) < c^*$; choose g_η as a KPP and concave function smaller than g satisfying $g'_\eta(0) = g'(0) - \eta$. Now just replace g by g_η in $(E_{\epsilon,\delta})$: this provides a subsolution with the same properties as z and completes the proof of the second part of Theorem 3.1. □

3.2. Initial data with compact support.

THEOREM 3.4. *We suppose that u_0 satisfies $u_0(x, y) = o(e^{\lambda x})$ as $x \to -\infty$, uniformly in y for each $\lambda < \lambda^*$ and $u_0(x, y) = o(e^{-\mu x})$ as $x \to +\infty$, uniformly in y for each $\mu < \lambda^{**}$, then u satisfies uniformly on every compact of $\bar{\Sigma}$.*

1. *$u(x - ct, y, t) \to 0$ when $t \to +\infty$ if $c \notin [c^{**}, c^*]$,*
2. *$u(x - ct, y, t) \to 1$ when $t \to +\infty$ if $c \in (c^{**}, c^*)$.*

The first assertion follows directly from Theorem 3.1. As a first step to the second assertion, we prove the following lemma, which was obtained independently by Lachand-Robert [14] in a different context.

LEMMA 3.5. *There exists $c' < c^*$ such that $\forall c \in (c', c^*)$, $\exists \lambda \in \mathbb{C} \setminus \mathbb{R}$ and $\varphi \in C^2(\bar{\omega}, \mathbb{C})$ with $\mathrm{Re}(\varphi) > 0$ in $\bar{\omega}$ which satisfies*

$$\begin{cases} -\Delta\varphi + (\lambda(\alpha(y) + c) - \lambda^2 - g'(0))\varphi = 0 & \text{in } \omega, \\ \partial_\nu\varphi = 0 \quad \text{on } \partial\omega. \end{cases}$$

Proof. Remember that for $\lambda \in \mathbb{R}$, we denote by $\mu_1(-\Delta + \lambda\beta_*)$ the first eigenvalue of the operator $-\Delta + \lambda(\alpha(y) + c^*)$; the eigenspace associated in $L^2(\omega)$ is one-dimensional. The family of operators $-\Delta + \lambda(\alpha(y) + c^*)$ depends analytically on λ, in the sense of Kato (see [16]). From the Kato–Rellich theorem, $\exists V$ a neighbourhood of λ^* in $\mathbb{C}$ such that there exists a simple eigenvalue $\tilde{\mu}_1(-\Delta + \lambda\beta_*)$ continuing μ_1 on all V analytically; there also exists a family of eigenfunctions ψ_λ in $L^2(\omega)$, analytic in λ.

Let us study the zeros of $F_c(\lambda) = \tilde{\mu}_1(-\Delta + \lambda\beta_*) + \lambda(c - c^*) - \lambda^2 - g'(0)$. F_c is analytic and converges locally uniformly to F_{c^*} as $c \to c^*$, λ^* being a zero of F_{c^*}, the Rouché theorem yields the existence, in a neighbourhood of c^*, of a family λ_c such that $F_c(\lambda_c) = 0$ and $\lambda_c \to \lambda^*$ as $c \to c^*$. We then have, for c near c^*: $\tilde{\mu}_1(-\Delta + \lambda_c\beta_*) + \lambda_c(c - c^*) - \lambda_c^2 - g'(0) = 0$; setting $\psi_c = \psi_{\lambda_c}$, we get

$$\begin{cases} -\Delta\psi_c + (\lambda_c(\alpha(y) + c) - \lambda_c^2 - g'(0))\psi_c = 0 & \text{in } \omega, \\ \partial_\nu\psi_c = 0 \quad \text{on } \partial\omega. \end{cases}$$

Notice that $\psi_c \to \psi_{c^*}$ as $c \to c^*$, in $L^2(\omega)$. Writing the equation satisfied by $\psi_c - \psi_{c^*}$, and using classical elliptic estimates, we get

$$||\psi_c - \psi_{c^*}||_{L^2(\omega)} \to 0 \Longrightarrow ||\psi_c - \psi_{c^*}||_\infty \to 0$$

as $c \to c^*$. Without loss of generality, we can assume that ψ_{c^*} is real and positive on $\bar{\omega}$ (since the eigenspace associated to λ^* is generated by a real positive function). We have that $\mathrm{Re}(\psi_c) \to \psi_{c^*}$ in $L^\infty(\omega)$, so $\mathrm{Re}(\psi_c) > 0$ for c near c^*. Now we need only prove that for $c \in \mathbb{R}$, with $c < c^*$ near enough, it is impossible to have $\lambda_c \in \mathbb{R}$; if it were possible, we could write $\mu_1(-\Delta + \lambda_c(\alpha(y) + c) - g'(0)) = \lambda_c^2$; but $c < c^*$ means that this never happens for $\lambda \in \mathbb{R}$ (see the definition of c^*). The lemma is proved with $\lambda = \lambda_c$ and $\varphi = \psi_c$. □

Now we prove Theorem 3.4. Let $\epsilon_0 > 0$ such that $c^{**} < c^* - \epsilon_0$. Let c_δ the c^* obtained when replacing $g'(0)$ by $g'(0) - \delta$, for some $\delta > 0$, in the principal eigenvalue problem. It is not difficult to see that c_δ increases to c^* as δ decreases to 0. So, for δ small enough, $c^* - \epsilon_0 < c_\delta < c^*$. We apply the lemma with $g'(0) - \delta$ instead of $g'(0)$ and c_δ instead of c^*. This provides, for $c < c_\delta$ near c_δ, (λ, φ) such that $\varphi_0 := \mathrm{Re}(e^{\lambda x}\varphi(y))$ is a real-valued solution of

$$\begin{cases} -\Delta\varphi_0 + (\alpha(y) + c)\partial_x\varphi_0 = (g'(0) - \delta)\varphi_0 & \text{in } \Sigma, \\ \partial_\nu\varphi_0 = 0 & \text{on } \partial\Sigma. \end{cases}$$

Since $\mathrm{Re}(\varphi) > 0$ in $\bar{\omega}$ and $\lambda \in \mathbb{C} \setminus \mathbb{R}$, there exist two functions in $C^1(\bar{\omega})$, $x_1(y)$ and $x_2(y)$, such that

$$\forall y \in \bar{\omega}, \ , x_1(y) < x_2(y) \quad \text{and} \quad \varphi_0(x_1(y), y) = \varphi_0(x_2(y), y) = 0,$$
$$\forall y \in \bar{\omega}, \quad \forall x \in (x_1(y), x_2(y)), \quad \varphi_0(x, y) > 0.$$

Set $D = \{(x, y) \in \Sigma; x_1(y) < x < x_2(y)\}$. Since g is C^1, there exists $\eta > 0$ such that $\forall s \in (0, \eta)$, $(g'(0) - \delta)s < g(s)$; so, without loss of generality, we can assume that $\varphi_0 \leq \eta$ on D, which implies

$$-\Delta\varphi_0 + (\alpha(y) + c)\partial_x\varphi_0 = (g'(0) - \delta)\varphi_0 \leq g(\varphi_0) \text{ in } D.$$

Now we follow the idea of Aronson and Weinberger in [2]. Solve the following problem:

$$\begin{cases} \partial_t w - \Delta w + (\alpha(y) + c)\partial_x w = g(w) & \text{in } \Sigma, \\ \partial_\nu w = 0 & \text{on } \partial\Sigma, \\ w(t = 0) = \varphi_0 \text{ in } D, \text{ and } 0 & \text{in } \Sigma - D. \end{cases}$$

Then $w(x, y, t) \geq \varphi_0(x, y)$ in D, $\forall t \geq 0$ (apply the maximum principle in D, taking into account the Neumann condition on $\partial D \cap \partial\Sigma$ and the Dirichlet condition on $\partial D \cap \Sigma$). Considering the function $m(x, y, t) = w(x, y, t + h) - w(x, y, t)$ for $h > 0$, we see that m is positive $\forall t \geq 0$ (maximum principle again), and so w is increasing in t. Then there exists a limit function $w_\infty(x, y)$ to which $w(x, y, t)$ converges as $t \to +\infty$. Using classical parabolic estimates and Ascoli theorem, we get the uniform convergence on every compact subset of $\bar{\Sigma}$, as well as for the derivatives $\partial_x w$, $\partial_t w$, and Δw. This implies

$$-\Delta w_\infty + (\alpha(y) + c)\partial_x w_\infty = g(w_\infty) \text{ in } \Sigma$$

with Neumann boundary conditions.

An argument similar to that of [5], Theorem 1 shows that w_∞ has to go to a constant limit as $|x| \to +\infty$, which can only be 0 or 1. But from the theory of Berestycki and Nirenberg, it follows that w_∞ is monotonic; since $c \in (c^{**}, c^*)$, it must be 0 or 1 identically, so it is 1 identically. To sum up, we have just constructed a function with compact support, with arbitrarily small supremum norm, and such that the solution of the corresponding Cauchy problem goes to 1 uniformly on every compact subset of $\bar{\Sigma}$. Applying the maximum principle, we generalize this result to any u_0 having compact support.

Now proceed similarly with $u(-x, y, t)$; we finally get that $u(x - ct, y, t) \to 1$ for some real $c \in (c^{**}, c^*)$ arbitrarily near c^{**}. The second assertion in Theorem 3.4 is then proved for reals $c \in (c^{**}, c^{**} + \epsilon) \cup (c^* - \epsilon, c^*)$ with any small $\epsilon > 0$. The general result follows from an application of the maximum principle as in Theorem 4.7; see §4. □

NB: For the proof of the second assertion in Theorem 3.4, we needn't suppose that $g(s)/s$ decreases.

4. Global stability of waves. In this section, we show that the travelling waves are stable in some sense, and give criteria for the creation of travelling fronts, as t grows to infinity.

In order to emphasize the main ideas, we first investigate particular cases, then generalize. Recall that g is still assumed to be a KPP term.

LEMMA 4.1. *Let $c \in \mathbb{R}$ and $\beta(y) = \alpha(y) + c$. Consider two solutions u and v of*

$$\begin{cases} \partial_t z - \Delta z + \beta(y)\partial_x z = g(z) & \text{in } \Sigma, \\ \partial_\nu z = 0 & \text{on } \partial\Sigma. \end{cases}$$

Also suppose that $u(t = 0) \leq v(t = 0)$, with values in $[0, 1]$ and that there exists $x_0 \in \mathbb{R}$ such that $u(x, y, 0) = v(x, y, 0)$, $\forall x \leq x_0$, $\forall y \in \omega$; then

$$0 \leq v(t) - u(t) \leq Ce^{\lambda^*[x-(c-c^*)t]} \text{ for some real } C.$$

Proof. The left-hand side follows from the maximum principle. For the right-hand side, we write for some w,

$$(\partial_t + A_c)(v - u) = g'(w)(v - u) \leq g'(0)(v - u)$$

since $v - u \geq 0$ and g is a KPP source term; then let then $e^{\lambda^* x}\varphi^*(y)$ be a positive exponential solution of

$$-\Delta z + (\alpha(y) + c^*)\partial_x z = g'(0)z$$

with the Neumann boundary condition on $\partial\Sigma$; then $h(x, y, t) = e^{\lambda^*[x-(c-c^*)t]}\varphi^*(y)$ satisfies $(\partial_t + A_c - g'(0))h = 0$ with the same boundary condition; we then choose a constant Λ such that $v(0) - u(0) \leq \Lambda h(0)$; the result then follows from the maximum principle. □

THEOREM 4.2. *Let $c > c^*$, $\beta(y) = \alpha(y) + c$, and consider $u(t)$ the solution of the problem*

$$(\mathcal{P}_c) : \begin{cases} \partial_t u - \Delta u + (\alpha(y) + c)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 & \text{on } \partial\Sigma, \\ u(t = 0) = u_0 & \text{in } \Sigma, \end{cases}$$

where u_0, with values in $[0,1]$, satisfies

$$u_0(x,y) \sim \phi_c(x,y) \text{ uniformly on } \bar{\omega} \text{ when } x \to -\infty,$$

then

$$u(x,y,t) \to \phi_c(x,y) \text{ when } t \to +\infty, \text{ uniformly on } (-\infty,a] \times \bar{\omega} \quad \forall a \in \mathbb{R}.$$

Proof. Let $a \in \mathbb{R}$ and $\epsilon > 0$. Set $\phi := \phi_c$. Let $\delta > 0$ such that

$$|\phi(x,y) - \phi(x-\delta,y)| < \epsilon/2 \quad \text{on } \Sigma$$

since $\phi_c(x,y) \sim e^{\lambda_c x}\varphi_c(y)$ as $x \to -\infty$, with $\lambda_c > 0$ and $\varphi_c > 0$ on $\bar{\omega}$, we have

$$\phi^\delta(x,y) := \phi(x-\delta,y) \sim e^{-\lambda_c \delta}\phi(x,y) \quad \text{when } x \to -\infty$$

so $\underline{u}_0 := \inf(u_0, \phi^\delta)$ satisfies $\underline{u}_0 \le \phi^\delta$ and $\underline{u}_0(x,y) = \phi^\delta(x,y)$ $\forall x \le x_0$ (for some real x_0); we denote by $\underline{u}$ the solution of the Cauchy problem with $\underline{u}(0) = \underline{u}_0$, and apply Lemma 4.1, which yields

$$0 \le \phi^\delta - \underline{u}(t) \le Ce^{\lambda^*[(x-(c-c^*)t]}.$$

Similar considerations on $\bar{u}_0 := \sup(u_0, \phi^{-\delta})$ provide

$$0 \le \bar{u}(t) - \phi^{-\delta} \le Ce^{\lambda^*[x-(c-c^*)t]}.$$

Now we just have to apply the maximum principle

$$\phi^\delta - Ce^{\lambda^*[x-(c-c^*)t]} \le \underline{u}(t) \le u(t) \le \bar{u}(t) \le \phi^{-\delta} + Ce^{\lambda^*[x-(c-c^*)t]} \qquad \square$$

Under an additional hypothesis, we now obtain uniform convergence in the whole cylinder.

THEOREM 4.3. *Let $c > c^*$, $\beta(y) = \alpha(y) + c$, and consider $u(t)$ the solution of the problem*

$$(\mathcal{P}_c): \begin{cases} \partial_t u - \Delta u + (\alpha(y)+c)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 \quad \text{on } \partial\Sigma, \\ u(t=0) = u_0 \quad \text{in } \Sigma, \end{cases}$$

with u_0 (taking its values in $[0,1]$) satisfying

1. *$u_0(x,y) \sim \phi_c(x,y)$ uniformly in ω as $x \to -\infty$*
2. *$\liminf_{x\to+\infty} u_0 > 0$ uniformly in ω, then*

$$\sup_{(x,y)\in\bar{\Sigma}} |u(x,y,t) - \phi_c(x,y)| \to 0 \text{ as } t \to +\infty.$$

We will make use of the following lemma. These kinds of sub- and supersolutions results have been widely used in one space dimension; see [10] and [21].

LEMMA 4.4. *Let $c > c^*$ and $\epsilon_0 \in (0,1)$, then $\exists s, K > 0$ such that $\psi(x,y,t) := h(t)\phi_c(x-m(t),y)$ satisfies $\partial_t\psi + A_c\psi - g(\psi) \le 0$ on Σ, $\forall t \ge 0$, with $h(t) := 1-\epsilon_0 e^{-st}$ and $m(t) := K(1-e^{-st})$.*

Proof. Let $u_1 \in (0,1)$ such that $\forall \epsilon_0 \in (0,1)$, $\exists T > 0$ satisfying

$$(1-\epsilon)g(u) - g((1-\epsilon)u) \le -T\epsilon u, \ \forall u \in [u_1,1], \ \forall \epsilon \in [0,\epsilon_0]$$

(such an u_1 exists as soon as $g'(0) > 0 > g'(1)$).

Set $N[\psi] = \partial_t \psi + A_c \psi - g(\psi)$, and $\varphi(x,y,t) = \phi_c(x - m(t), y)$; we compute

$$N[\psi] = s\epsilon_0 e^{-st}\varphi - hsK\partial_x\varphi + hg(\varphi) - g(h\varphi).$$

There exists X_0 such that $\forall x \geq X_0$, $\forall y \in \omega$, $\phi_c(x,y) \geq u_1$, so if $x - m(t) \geq X_0$, we have $hg(\varphi) - g(h\varphi) \leq -T\epsilon_0 e^{-st}\varphi$ and then $N[\psi] \leq e^{-st}\varphi(s-T)\epsilon_0$; choosing $s = T$, we have $N[\psi] \leq 0$; if $x - m(t) \leq X_0$, notice that there exists a constant $C > 0$ such that $\partial_x\varphi/\varphi \geq C$, because of the asymptotic behaviour of ϕ_c as $x \to -\infty$. Write

$$N[\psi] \leq s\epsilon_0 e^{-st}\varphi - hsCKe^{-st}\varphi + g(\varphi) - g(h\varphi),$$

$$N[\psi] \leq e^{-st}\varphi(s\epsilon_0 - hsCK + C'\epsilon_0).$$

Now choose $K = (s + C')\epsilon_0/(1 - \epsilon_0)sC$, which yields $N[\psi] \leq 0$ as well. □

Here is the proof of Theorem 4.3. In the case where $u_0(x,y) > 0$ in $\bar{\Sigma}$, the hypothesis guarantees there exists some $h_0 \in (0,1)$ and some $x_0 \in \mathbb{R}$, satisfying $h_0\phi_c(x - x_0, y) \leq u_0(x,y)$ in Σ. Apply Lemma 4.4 with $\epsilon_0 = 1 - h_0$. Then $\psi(x,y,t) \leq u(x,y,t)$, $\forall t \geq 0$; this, together with Theorem 4.2, proves that the convergence is uniform on $\bar{\Sigma}$.

If we do not assume $u_0(x,y) > 0$ in $\bar{\Sigma}$, we still have, by the maximum principle, that $u(x,y,t_0) > 0$ as soon as $t_0 > 0$. Notice that condition 1 is preserved at all positive times (use the same arguments as in the proof of Theorem 4.2); condition 2 is in fact also preserved: just compare u_0 to some C^∞ and the increasing function $h_0(x,y)$, which is 0 if $x < M$ and ϵ if $x > M+1$; solve the corresponding Cauchy problem, which gives some $h(x,y,t)$ increasing in x, so that obviously $\liminf_{x\to+\infty} h(x,y,t) > 0$, for each positive t.

Both conditions being preserved,we use the semigroup property and prove Theorem 4.3 in the general case. □

Now we give a more powerful version of this result.

THEOREM 4.5. *Let $c > c^*$, $\beta(y) = \alpha(y) + c$, and consider $u(t)$ the solution of the problem*

$$(\mathcal{P}_c): \begin{cases} \partial_t u - \Delta u + (\alpha(y) + c)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 & \text{on } \partial\Sigma, \\ u(t=0) = u_0 & \text{in } \Sigma, \end{cases}$$

with u_0 (taking its values in $[0,1]$) satisfying

1. *$u_0(x,y) \sim a(y)e^{\lambda_c x}$ uniformly on ω as $x \to -\infty$, where $a(y)$ is continuous and not identically zero;*
2. *$\liminf_{x\to+\infty} u_0 > 0$ uniformly on ω,*

then

$$\exists x_0 \in \mathbb{R} \text{ such that } \sup_{(x,y)\in\bar{\Sigma}} |u(x,y,t) - \phi_c(x + x_0, y)| \to 0 \text{ as } t \to +\infty.$$

First, let us state a lemma.

LEMMA 4.6. *Let u be a solution of $(\mathcal{P}_c)$, with $c > c^*$; suppose $u(x,y,0)$ satisfies the first condition in Theorem 4.5. Let $\psi(y,t)$ be the solution of*

$$\begin{cases} \partial_t\psi - \Delta\psi + (\lambda_c\beta(y) - \lambda_c^2)\psi = g'(0)\psi & \text{in } \omega, \\ \partial_\nu\psi = 0 & \text{on } \partial\omega, \\ \psi(y,0) = a(y), \end{cases}$$

then

$$u(x,y,t) \sim \psi(y,t)e^{\lambda_c x} \text{ as } x \to -\infty, \forall t \geq 0, \text{ uniformly in } \omega.$$

Proof. Set $u(x,y,t) = e^{\lambda_c x}v(x,y,t)$. Notice that $v(x,y,t)$ is bounded on every $[0,T]$. To prove it, notice that the hypothesis on u_0 implies the existence of a wave ϕ_c and a real x_0 such that $(u_0 - \phi_c)(x,y) < 0$ for all $x < x_0$; then apply Lemma 4.1:

$$u(x,y,t) < \phi_c(x,y) + Ce^{\lambda^*[x-(c-c^*)t]} \quad \forall t > 0$$

so

$$v(t) < e^{-\lambda_c x}\phi_c(x,y) + Ce^{(\lambda^*-\lambda)x-\lambda^*(c-c^*)t}$$

since $\lambda_c < \lambda^*$, this yields an upper bound for $v(t)$ on $\mathbb{R}^- \times \omega$, $\forall t \in [0,T]$, whereas $v(t)$ is clearly bounded on $\mathbb{R}^+ \times \omega$ (remember that $u \leq 1$).

For $\theta \in \mathbb{R}$, set $w(x,y,t) = v(x,y,t) - \theta\psi(y,t)$ and write the partial differential equation satisfied by w:

$$Lw := \partial_t w - \Delta w + \beta(y)\partial_x w + (\lambda_c\beta(y) - \lambda_c^2 - g'(0))w = e^{-\lambda_c x}g(e^{\lambda_c x}v) - g'(0)v,$$

with g being C^2, we can write $g(s) = g'(0)s + (s^2/2)g''(\eta s)$, so $|Lw| \leq Ke^{-\lambda_c x}e^{2\lambda_c x}v^2$, but we know that v is bounded on each $[0,T]$, therefore

$$|Lw| \leq K(T)e^{\lambda_c x} \quad \forall t \in [0,T].$$

Let M be real, and $h(x,t) = e^{\lambda_c x + Mt}$; choose $T > 0$. We have $Lh = (M + 2(\lambda_c\beta(y) - \lambda_c^2) - g'(0))h$, so for some large enough $M > 0$, we have

$$Lh \geq K(T)h = K(T)e^{\lambda_c x + Mt} \geq K(T)e^{\lambda_c x}.$$

If $\theta > 1$, we have $w(x,y,0) < 0$ for $x < x(\theta)$ (a real depending on θ), and since $w(x,y,0)$ is bounded in Σ, we get $w(x,y,0) < Nh(x,0)$ for N large enough. The maximum principle then implies

$$w(x,y,t) < Nh(x,t) \quad \forall t \in [0,T],$$

$$v(x,y,t) < \theta\psi(y,t) + Ne^{\lambda_c x + Mt}.$$

Meanwhile, $a(y)$ being nonnegative and not identically zero, we know that $\psi(y,t) \geq \epsilon(T) > 0$ for all $t \in [1/T, T]$, so

$$v(x,y,t) < (\theta + N(T)e^{\lambda_c x})\psi(y,t) \quad \forall t \in [1/T, T],$$

then $\limsup_{x\to-\infty}(v(x,y,t)/\psi(y,t)) \leq \theta$, uniformly in $\omega \times [1/T, T]$; the same treatment for $\theta < 1$ provides $\liminf_{x\to-\infty}(v(x,y,t)/\psi(y,t)) \geq \theta$, uniformly in $\omega \times [1/T, T]$, which ends the proof of Lemma 4.6. □

We keep the notation of Lemma 4.6 for the proof of Theorem 4.5: Consider the operator $T = -\Delta + (\lambda_c\beta(y) - \lambda_c^2 - g'(0))$, with Neumann boundary condition; T is self-adjoint with 0 as first eigenvalue; then it is classical to see that, in $L^2(\omega)$, $\psi(y,t) \to \varphi_c(y)$, as $t \to +\infty$, where $\varphi_c(y)$ is the only positive eigenfunction satisfying $\int_\omega a\varphi_c = \int_\omega \varphi_c^2$.

Once again, we use the classical parabolic estimates to prove that the convergence happens in $L^\infty(\omega)$.

Let $\epsilon > 0$, so there exists t_0 such that $\forall t \geq t_0$, $(1-\epsilon)\varphi(y) < \psi(y,t) < (1+\epsilon)\varphi(y)$. Let ϕ_c and ϕ'_c the travelling waves, with speed c, satisfying $\phi_c \sim (1-\epsilon)\varphi(y)e^{\lambda_c x}$ and $\phi'_c \sim (1+\epsilon)\varphi(y)e^{\lambda_c x}$ as $x \to -\infty$. We apply Theorem 4.3 with initial data $\inf\{\phi_c(x,y); u(x,y,t_0)\}$ and $\sup\{\phi'_c(x,y); u(x,y,t_0)\}$; $u(x,y,t)$ stays between the solutions of the corresponding Cauchy problems (maximum principle), but these converge uniformly on $\bar{\Sigma}$ to ϕ_c and ϕ'_c, respectively. Just notice now that $||\phi_c - \phi'_c||_{L^\infty(\Sigma)}$ is arbitrarily small when ϵ is. This proves Theorem 4.5. □

Keeping the notation of Theorem 4.5, we denote by ϕ_c the only travelling front such that $\phi_c(x,y) \sim e^{\lambda_c x}\varphi_c(y)$ as $x \to -\infty$. We have an explicit expression for x_0:

$$x_0 = \frac{1}{\lambda_c}\left(\mathrm{Log}\int_\omega a\varphi_c - \mathrm{Log}\int_\omega \varphi_c^2\right).$$

Our last result concerns the creation of fronts at both ends of the cylinder.

THEOREM 4.7. *Let $u(x,y,t)$ be a solution of $(\mathcal{P})$; suppose there exist $\lambda \in (0,\lambda^*)$ and $\tilde{\lambda} \in (0,\lambda^{**})$ satisfying*

$$u_0(x,y) \sim a(y)e^{\lambda x} \text{ near } -\infty \text{ uniformly in } \omega,$$
$$u_0(x,y) \sim b(y)e^{-\tilde{\lambda}x} \text{ near } +\infty \text{ uniformly in } \omega$$

for some continuous, nonidentically zero $a(y)$ and $b(y)$; then if we denote by c and c' the wave speeds associated to λ and $\tilde{\lambda}$, respectively, with $c' < c^{**} < c^* < c$, we have

$$\sup_{(x,y)\in\Sigma} |u(x,y,t) - \inf\{\phi_c(x+ct,y), \tilde{\phi}_c(x+c't,y)\}| \to 0 \quad \text{as } t \to +\infty.$$

Proof. Using Lemma 4.6, and proceeding as in the proof of Theorem 4.5, we have that $u(x-ct,y,t) \to \phi_c(x,y)$, uniformly on each $(-\infty,a]\times\omega$ and $u(x-c't,y,t) \to \tilde{\phi}_{c'}(x,y)$, uniformly on each $[a,+\infty)\times\omega$.

Let $\epsilon > 0$; there exist X_1 and X_2 such that $\phi_c(x,y) \geq (1-\epsilon)$, $\forall x \geq X_1$, $\forall y \in \omega$, and $\tilde{\phi}_{c'}(x,y) \geq (1-\epsilon)$, $\forall x \leq X_2$, $\forall y \in \omega$. Set $D = \{(x,y,t)/X_1 - ct < x < X_2 - c't\}$; choose some large t_0 satisfying $t \geq t_0 \Longrightarrow u(X_1 - ct, y, t) \geq \phi_c(X_1, y) - \epsilon \geq 1 - 2\epsilon$ and $u(X_2 - c't, y, t) \geq \tilde{\phi}_{c'}(X_2, y) - \epsilon \geq 1 - 2\epsilon$. Set $g_\epsilon(s) = ms(1 - 2\epsilon - s)$ with m small enough to have $g_\epsilon \leq g$; we are going to apply the maximum principle on $D \cap \{t \geq t_0\}$: Let $h(t)$ satisfy $\dot{h} = g_\epsilon(h)$ and $h(0) = \inf\{1-2\epsilon; \inf_{D\cap\{t=t_0\}} u(x,y,t_0)\}$; now write

$$(\partial_t + A)(u - h) = g(u) - g_\epsilon(h) \geq g(u) - g(h).$$

Considering the boundary conditions, this shows that $u(t) \geq h(t+t_0)$ in $D \cap \{t = t_0\}$.

Therefore, there exists t_1 such that $\forall t \geq t_1$, $u(x,y,t) \geq 1 - 3\epsilon$ in D. This ends the proof of Theorem 4.7. □

5. Local stability in weighted spaces. In this section, we adopt another point of view, based upon the spectral analysis of operator A_c in a suitable weighted space. This kind of analysis appeared for the first time in the work of Sattinger [19], with a KPP source term. This section generalises [19]; furthermore, we show how to improve very easily the uniform convergence results of the last section in the case of a KPP source term, when the initial datum has a suitable behaviour as $x \to +\infty$. The general case is more intricate and is presented in the next section.

5.1. Framework. We linearise problem $(\mathcal{P}_c)$ near ϕ_c by setting $u_0 = \phi_c + v_0$ and $u(t) = \phi_c + v(t)$, $\forall t > 0$, with $c > c^*$ fixed once and for all. We obtain

$$\partial_t u + A_c u = g(u) \Leftrightarrow \partial_t v + A_c v = g(\phi_c + v) - g(\phi_c).$$

Set $L = A_c - g'(\phi_c)$ and $f(v) = g(\phi_c + v) - g(\phi_c) - vg'(\phi_c)$, which leads to the formulation

$$\begin{cases} \partial_t v + Lv = f(v) & \text{in } \Sigma, \\ \partial_\nu v = 0 & \text{on } \partial\Sigma, \\ v(0) = v_0. \end{cases}$$

Choice of a space. Set $\beta_\eta = \beta + \eta - \bar{\beta}$ ($\bar{\beta}$ denotes the average of β); for $0 < \eta < \bar{\beta}$ sufficiently close to $\bar{\beta}$, the following problem

$$(E_\eta) : \begin{cases} -\Delta_y \varphi - g'(0)\varphi + \lambda \beta_\eta \varphi = \lambda^2 \varphi & \text{in } \omega, \\ \partial_\nu \varphi = 0 & \text{on } \partial\omega, \end{cases}$$

admits exactly two positive principal eigenvalues; we denote by $r = r(\eta)$ the smallest one, and consider an associated eigenfunction ψ, which is positive on ω.

Remark 1. $r > \lambda_c$ and $r(\eta) \searrow \lambda_c$ as $\eta \nearrow \bar{\beta}$.

Let $\Gamma \in C^\infty(\mathbb{R})$ such that $0 \leq \Gamma \leq 1$, $\Gamma = 0$ on $(-\infty, 0]$ and $\Gamma = 1$ on $[1, +\infty)$; let

$$w_1(x,y) = (1 - \Gamma(x))e^{rx}\psi(y) + \Gamma(x+1) \text{ and } w(x,y) = 1/w_1(x,y)$$

set

$$X = \{u \in C_0(\bar{\Sigma}); wu \in C_0(\bar{\Sigma})\},$$

where $C_0(\bar{\Sigma})$ is the set of all bounded uniformly continuous functions on $\bar{\Sigma}$, which tend to 0 as $|x| \to \infty$, uniformly on $\bar{\omega}$.

We equip X with the norm $||u||_X = ||wu||_\infty$, so that

$$M : X \longrightarrow C^0(\bar{\Sigma}),$$

$$u \mapsto wu,$$

is an isometric isomorphism between $(X, ||.||_X)$ and $(C^0(\bar{\Sigma}), ||.||_\infty)$.

Remark 2. $\partial_x \phi_c \notin X$ This is clear from the behaviour of $\partial_x \phi_c$ as $x \to -\infty$.

Transformation of operator L. To the operator L defined on X, we associate $\tilde{L} = MLM^{-1}$ defined in $C^0(\bar{\Sigma})$, in other words $\forall v \in C^0(\bar{\Sigma})$ smooth, $\tilde{L}v = wL(v/w)$; recall that $L = -\Delta + \beta(y)\partial_x - g'(\phi_c)$, therefore we get $\tilde{L}$

$$\tilde{L}v = -\Delta v + B(x,y).\nabla v + C(x,y)v,$$

where

$$B(x,y).\nabla v = \beta(y)\partial_x v + 2\frac{\nabla w}{w}.\nabla v$$

and

$$C(x,y) = \frac{\Delta w}{w} - 2\frac{|\nabla w|^2}{w^2} - \beta(y)\frac{\partial_x w}{w} - g'(\phi_c).$$

Thanks to our choice for w, we get that $B(x,y)$ is bounded, and

$$C(x,y) = \begin{cases} r(\bar{\beta}-\eta) + g'(0) - g'(\phi_c) & \text{for } x < -1, \\ -g'(\phi_c) & \text{for } x > 1. \end{cases}$$

Set $\gamma(x) = r(\bar{\beta}-\eta)(1-\Gamma(x)) - g'(1)\Gamma(x+1)$; notice that $\inf_{I\!R}\gamma > 0$ and that $|C(x,y)-\gamma(x)| \to 0$ as $|x| \to \infty$, uniformly on $\bar{\omega}$; set

$$\tilde{T}v = -\Delta v + B(x,y).\nabla v + \gamma(x)v,$$

$$\tilde{S}v = [C(x,y)-\gamma(x)]v,$$

write $\tilde{L} = \tilde{S} + \tilde{T}$ and $L = S + T$ with $S = M^{-1}\tilde{S}M$ and $T = M^{-1}\tilde{T}M$; finally, define the domains

$$D(\tilde{L}) = D(\tilde{T}) = \left\{ u \in C_0(\bar{\Sigma}) \cap \left(\bigcap_{p\geq 1} W^{2,p}_{loc}(\Sigma) \right) ; \Delta u \in C^0(\bar{\Sigma}) \text{ and } \partial_\nu u = 0 \text{ on } \partial\Sigma \right\},$$

$$D(L) = M^{-1}(D(\tilde{L})) = \left\{ u \in X \cap \left(\bigcap_{p\geq 1} W^{2,p}_{loc}(\Sigma) \right) ; \Delta u \in X \text{ and } \partial_\nu u = 0 \text{ on } \partial\Sigma \right\}.$$

LEMMA 5.1. *$D(L)$ is dense in X.*

5.2. Spectral properties of L.

LEMMA 5.2. *L is sectorial in X, more precisely $\exists a < 0$, $\gamma_0 > 0$, and $\alpha \in [0, \frac{\pi}{2})$ such that*

1. *$\sigma(L)$ (the spectrum of L in X) is contained in the cone*

$$C_{a,\alpha} = \{ z \in \mathbb{C} \, ; |\mathrm{Arg}(z-a)| < \alpha \}.$$

2. *If $\lambda \in \sigma(L)$ and $\mathrm{Re}(\lambda) < \gamma_0$, then λ is an eigenvalue for L.*

The proof is similar to the one given in [17, Thm. 4.1].

THEOREM 5.3. *If λ is an eigenvalue for L in X, then either $\lambda = 0$ or $\mathrm{Re}(\lambda) > 0$.*

Proof. It is nearly the proof of Theorem 1 in [6], and in fact much simpler since there is nothing to prove as $x \to -\infty$.

Let $u \in X$ such that $Lu = \lambda u$; differentiating $A_c\phi_c = g(\phi_c)$ yields $L(\partial_x\phi_c) = 0$; with u being in X, we immediately have

$$\frac{u}{\partial_x\phi_c} \to 0 \quad \text{when } x \to -\infty;$$

the only point is to prove the same as $x \to +\infty$. Indeed, if we denote by λ^+ the exponent which rules the behaviour of ϕ_c at $+\infty$, and then if $\mathrm{Re}(\lambda) \leq 0$ and $\lambda \neq 0$, there exists $\epsilon > 0$ such that $|u(x,y)| \leq e^{-(\lambda^+ + \epsilon)x}$ in Σ. (See Lemmas 5.1, 5.2, 5.3, and 5.6 in [6].)

Finally, the maximum principle applied to the function $\mathrm{Re}(e^{-\lambda t}u(x,y))/\partial_x\phi_c$ shows that we have $u = 0$. □

PROPOSITION 5.4. *$L : D(L) \longrightarrow X$ is one-to-one.*

Proof. Let $u \in D(L)$ such that $Lu = 0$; from [7] and Lemma 2.5, there exist α_-, α_+, such that

$$\frac{u}{\phi_x} \to \alpha_- \text{ as } x \to -\infty,$$
$$\frac{u}{\phi_x} \to \alpha_+ \text{ as } x \to +\infty.$$

Now proceed as in [17], Proposition 3.4: This yields (u/ϕ_x) =constant. Now since u is supposed to decay faster than ϕ_x, $u = 0$. □

Remark 3. This proves that $Lu = 0$, $u \in C_0(\bar{\Sigma})$ implies that u is proportional to $\partial_x \phi_c$.

5.3. Conclusion: Local stability and exponential convergence.

PROPOSITION 5.5. *L is a sectorial operator in X with spectrum $\sigma(L)$ contained in a cone $C_{a,\alpha}$ with $a > 0$ and $\alpha \in [0, \frac{\pi}{2})$.*

In particular, $\mathrm{Re}(\sigma(L)) > a > 0$.

This is an obvious consequence of the above paragraph.

THEOREM 5.6. *$\exists \omega > 0$, $\rho > 0$, and $M \geq 0$ such that, if $||v_0||_X \leq \rho/2M$, then the problem*

$$\begin{cases} dv/dt + Lv = f(v) & \forall t > 0 \\ v(0) = v_0 \end{cases}$$

admits exactly one solution v defined on $(0, +\infty)$ and satisfying $||v(t)||_X \leq 2Me^{-\omega t}||v_0||_X$.

Proof. Just use Theorem 5.1.1 in [11]. □

Now we may use the global stability theorems of §4 to get an exponential global stability result. Assume for simplicity that g is concave.

PROPOSITION 5.7. *Let u_0 be uniformly continuous in $\bar{\Sigma}$ such that $u_0(x, y) = \phi_c(x, y)(1 + O(e^{rx}))$ as $x \to -\infty$, for some positive real r, and $\liminf_{x \to +\infty} u_0 > 0$. Then*

$$||u(t) - \phi_c||_\infty \leq Ke^{-\kappa t},$$

where K and κ are two positive real numbers.

Proof. Let $w(x, y)$ be a positive exponential solution of

$$\begin{cases} -\Delta w + \beta_\eta \partial_x w = g'(0)w & \text{in } \Sigma, \\ \partial_\nu w = 0 & \text{on } \partial\Sigma, \end{cases}$$

with η close enough to $\bar{\beta}$ so that

$$w(x, y)/e^{(\lambda_c + r)x} \to +\infty \text{ as } x \to -\infty.$$

We have

$$\begin{aligned}(-\Delta + \beta\partial_x)(\phi_c + Cw) &= g(\phi) + \eta C \partial_x w + g'(0)Cw \\ &\geq g(\phi) + g'(0)Cw \geq g(\phi + Cw),\end{aligned}$$

for each constant C (here we have used the concavity of g). Symmetrically, we have

$$\begin{aligned}(-\Delta + \beta\partial_x)(\phi_c - Cw) &= g(\phi) - \eta C \partial_x w - g'(0)Cw \\ &\leq g(\phi) - g'(0)Cw \leq g(\phi - Cw).\end{aligned}$$

We infer from the maximum principle that, for C large enough, the following inequalities hold:

$$\phi_c - Cw \leq u(t) \leq \phi_c + Cw.$$

Now we use the global stability theorem and then we are in a position to apply the local exponential stability theorem. □

6. Extensions to non-KPP source terms. We drop here the KPP assumption. The global stability results for the fast waves—i.e., $c > c_*$—are the same as in the preceeding sections, but the proofs are more technical and appeal to some ideas presented in [18]. New stability results are derived when an additional assumption is made on g.

6.1. Global stability for rapid waves. We want to prove a result similar to Theorem 4.5; let us recall it.

THEOREM 6.1. *Let $c > c^*$, $\beta(y) = \alpha(y) + c$, and consider $u(t)$ the solution of problem*

$$(\mathcal{P}_C) : \begin{cases} \partial_t u - \Delta u + (\alpha(y) + c)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 \quad \text{on } \partial\Sigma, \\ u(t=0) = u_0 \quad \text{in } \Sigma, \end{cases}$$

with u_0 (taking its values in $[0,1]$) satisfying

1. *$u_0(x,y) \sim a(y)e^{\lambda_c x}$ uniformly on ω as $x \to -\infty$, where $a(y)$ is continuous and not identically zero;*
2. *$\liminf_{x\to+\infty} u_0 > 0$ uniformly on ω, then*

$$\exists x_0 \in \mathbb{R} \text{ such that } \sup_{(x,y)\in\bar{\Sigma}} |u(x,y,t) - \phi_c(x + x_0, y)| \to 0 \quad \text{as } t \to +\infty.$$

We need only prove this when $a(y) = \varphi_c(y)$ and $u_0(x,y) = \varphi_c(y)e^{\lambda_c x}(1 + O(e^{rx}))$. To see this, one may proceed as in [18, Lemma 2.3]; this is the way we will follow.

Before we start the proofs, notice that, just as in the proof of Theorem 4.5, we can obtain an explicit expression for the asymptotic shift x_0.

Let $\Gamma \in C^\infty(\mathbb{R})$ such that $0 \leq \Gamma \leq 1$, $\Gamma = 0$ on $(-\infty, 0]$ and $\Gamma = 1$ on $[1, +\infty)$.

LEMMA 6.2. *There exist three positive real M, q_0, and s, such that*

$$u(x,y,t) \leq \phi_c(x + M, y) + q_0 e^{-st}\Gamma(x + M).$$

Proof. Just as in [18, Lemma 5.1]. □

Next we need an even sharper control on the decay of the solution $u(x,y,t)$ as $x \to -\infty$.

LEMMA 6.3. *$\forall r' > r$, $\exists C > 0$ such that, $\forall t \geq 0$,*

$$|u(x,y,t) - \phi_c(x,y)| \leq Ce^{(\lambda_c + r')x}.$$

Proof. Let δ and x_0 such that

- $g'(0) < \delta$;
- the equation $\mu_1(-\Delta + \lambda\beta - \lambda^2 - \delta) = 0$ admits solutions; let $w(x,y)$ be a corresponding exponential solution;

- $\forall (x,y,t) \in (-\infty, x_0] \times \omega \times \mathbb{R}^+$, we have

$$\frac{g(u(x,y,t)) - g(\phi_c(x,y))}{u(x,y,t) - \phi_c(x,y)} \leq \delta.$$

Set $v = u - \phi_c$. The function w is an upper solution for the equation on v. So, for C large enough, we have $Cw \geq u - \phi_c$. The same method works if we look for a subsolution; just consider $\phi_c - u$. □

PROPOSITION 6.4. *There exist three positive real numbers M, q_0,and s, such that*

$$u(x,y,t) \geq \phi_c(x - M, y) - q_0 e^{-st}.$$

Proof. Let $\epsilon > 0$ such that the ball with radius ϵ and center ϕ_c (in the norm $1 + e^{-(\lambda_c + r/2)x}$), be in the attraction domain of ϕ_c. Choose $t_0 > 0$ such that $\forall t \geq t_0$, $\lim_{x \to +\infty} u(x,y,t) \geq 1 - \epsilon/2$. We can construct $\underline{u}_0(x,y)$ satisfying the following properties:

- $\underline{u}_0(x,y) = \phi_c(x,y)(1 + e^{2rx/3}(1 + o(1)))$ for $x \leq -M$;
- $\liminf_{x \to +\infty} \underline{u}_0(x,y) \geq 1 - \epsilon/2$;
- $\sup_{(x,y) \in \Sigma} |\underline{u}_0(x,y) - \phi_c(x,y)|.(1 + e^{-(\lambda_c + r/2)x}) \leq \epsilon$.

The solution of the Cauchy problem, with initial datum $\underline{u}_0(x - M, y)$, will converge exponentially to $\phi_c(x - M, y)$. Moreover, for large enough M, $\underline{u}_0(x,y) \leq u_0(x,y)$. □

Now we prove Theorem 6.1. Let

$$X_{\delta,r} = \{u_0 \in \mathrm{UC}(\bar{\Sigma}); u_0 \nearrow, \lim_{x \to +\infty} u_0(x,y) \geq \delta, \text{ and } |u_0(x,y) - \phi_c(x,y)| \leq C e^{(\lambda_c + 2r/3)x}\},$$

where UC is the set of bounded uniformly continuous functions. Also set

$$C_{\delta,r} = \{u_0 \in X; S(t)u_0 \to \phi_c \text{ in the weighted norm } 1 + e^{-(\lambda_c + r/2)x}\}.$$

It is clear from the local stability theorems that $C_{\delta,r}$ is open in $X_{\delta,r}$; in order to prove that it is also closed, we state the following proposition.

PROPOSITION 6.5. *Let $u_{10} \in X_{\delta,r}$. There exists ε_0 such that, for every $\varepsilon < \varepsilon_0$,*

$$\forall u_{20} \in X_{\delta,r}, \quad ||u_{10} - u_{20}||_{X_{\delta,r}} \leq \epsilon \implies \sup_{(x,y) \in \Sigma} (1 + e^{-\lambda_c x/2})|S(t)u_{10} - S(t)u_{20}| \leq C\epsilon.$$

Proof. Just do as in [18, Prop. 5.3]. The proof involves the construction of upper and lower solutions, and the use of parabolic Harnack inequlities "up to the boundary." □

COROLLARY 6.6. *If $u_0 \in X_{\delta,r}$, then $S(t)u_0$ converges towards ϕ_c.*

To end the proof of the theorem, just notice that it is always possible to find $\underline{u}_0 \in X_{\delta,r}$ and $\bar{u}_0 \in X_{\delta',r'}$, such that $\underline{u}_0 \leq u_0 \leq \bar{u}_0$. □

6.2. Stability of the wave with speed c^*. Here we suppose that $c^* > c_0$. Since the results that are stated can be proved by almost the same methods as in the above sections, we will only give the main lines of the proofs, leaving it to the interested reader to check the details.

We first show that the case $c^* > c_0$ is indeed a possible one, and proceed as in as Berestycki and Nirenberg in [8]. We multiply by $\partial_x \phi$ the equality

$$-\Delta\phi + (c^* + \alpha(y))\partial_x \phi = g(\phi)$$

and we integrate on Σ, which yields

$$\int_\Sigma (c^* + \alpha(y))[\partial_x \phi]^2 dxdy = G(1)$$

where $G(s) = \int_0^s g(\sigma)d\sigma$; the multiplication by $1 - \phi$ yields

$$(c^* + < \alpha >) \geq \frac{2}{|\omega|} \int_\Sigma |\nabla \phi|^2 dxdy,$$

$$(c^* + \sup \alpha)^2 \geq \frac{2}{|\omega|} \int_\Sigma |\nabla \phi|^2 (c^* + \sup \alpha) dxdy$$

so we get $(c^* + \sup \alpha)^2 \geq \frac{2}{|\omega|} G(1)$, that is to say

$$c^* \geq \sqrt{\frac{2}{|\omega|} G(1)} - \sup \alpha.$$

Now remember that $c_0(g'(0)) \to 0$ as $g'(0) \to 0$; then we just have to choose $g'(0)$ small and $G(1)$ large enough. □

As for the asymptotic behaviour of ϕ_{c^*} as $x \to -\infty$, we have to take into account the case when ϕ_{c^*} decays at the usual rate, and the case when ϕ_{c^*} decays at the accidental rate.

Case 1. If ϕ_{c^*} decreases at the usual rate, all the results for rapid waves are still available for ϕ_{c^*}. We leave it to the interested reader to check the analogue of Theorem 6.1 for ϕ_{c^*}.

Case 2. If ϕ_{c_*} decreases at the accidental rate, we get the same results as in the "ignition temperature" case: Local stability in weighted spaces with small exponents, and convergence to travelling waves for initial data that are increasing in x (see [18]). Namely, the weight functions that are suitable in this case are not sufficient to kill the translation invariance. Here is the precise result.

THEOREM 6.7. *Let $u(t)$ be the solution of*

$$(\mathcal{P}_*) : \begin{cases} \partial_t u - \Delta u + (\alpha(y) + c_*)\partial_x u = g(u) & \text{in } \Sigma, \\ \partial_\nu u = 0 \quad \text{on } \partial\Sigma, \\ u(t=0) = u_0 \quad \text{in } \Sigma. \end{cases}$$

Assume moreover that the initial datum u_0 satisfies

1. $u_0(x,y) = O(e^{rx})$ *as* $x \to -\infty$, *with* $r > 0$;
2. $\liminf_{x \to +\infty} u_0 > 0$ *uniformly on* ω.

Then the following is true.

1. *If u_0 has the form $u_0(x,y) = \phi_{c_*}(x,y) + \varepsilon v_0(x,y)$ (with v_0 decaying as e^{rx} as $x \to -\infty$), then for $\varepsilon_0 > 0$ small enough there exists $\gamma(\varepsilon) \in C^1(-\varepsilon_0, \varepsilon_0)$ and*

$$\sup_{(x,y)\in\bar{\Sigma}} |u(x,y,t) - \phi_c(x + \varepsilon\gamma(\varepsilon), y)| = O(e^{-\omega t}) \text{ as } t \to +\infty.$$

2. *If u_0 takes its values in $(0,1)$ is increasing in the x direction then $\exists x_0 \in \mathbb{R}$ and $\omega > 0$ such that*

$$\sup_{(x,y)\in\bar{\Sigma}} |u(x,y,t) - \phi_c(x + x_0, y)| = O(e^{-\omega t}) \text{ as } t \to +\infty.$$

Proof. Point 1 may be proved by the same method as in §5 as far as the linear stability is concerned. Nonlinear stability results are then proved in weighted spaces with exponents $< \lambda^a_{c_*}$; therefore one must go through all the steps of the proof of Theorem 1 in [17].

Let us now give the main lines of the proof of 2. Let $X_{r,\delta}$ and $C_{r,\delta}$ be defined as in the above subsection; we want to prove that $C_{r,\delta}$ is closed in $X_{r,\delta}$. One way to do it is as follows.

1. Prove an inequality of the type

$$\phi_{c_*}(x - M, y) - q_1 e^{-st} \leq u(x, y, t) \leq \phi_{c_*}(x + M, y) + q_2 e^{-st}.$$

This may be done as in Proposition 6.4.

2. Prove a uniform estimate as in Proposition 6.5. □

Remark 4. This method yields an extension to the second author's Theorem 1.2 in [18] to initial data having an arbitrary exponential decay as $x \to -\infty$.

Remark 5. The assumption $c_* > c_0$ is less artificial than it seems to be. Indeed, in combustion theory, in the framework of large activation energies, the so-called ZFK source term satisfies this assumption; see [4]. As such, the corresponding model should share many features with the "ignition temperature" model, investigated in [18]. This is exactly what we have just proved.

Acknowledgments. It is our pleasure to thank Prof. H. Berestycki for the kind interest that he always gave to our work, for his helpful advice, and for his useful remarks.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, Comm. Pure Appl. Math., 16 (1959), pp. 623–727; 17 (1964), pp. 35–92.

[2] D. G. Aronson and H. F. Weinberger *Nonlinear diffusion in population genetics, combustion and nerve propagation*, in Partial Differential Equations and Related Topics, Lecture Notes in Math. 446, Springer-Verlag, New York, 1975, pp. 5–49.

[3] ———, *Multidimensional diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–58.

[4] H. Berestycki and B. Larrouturou, *Quelques aspects mathematiques de la propagation des flammes prémélangées*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, 10, Brezis and Lions, eds., Pitman–Longman, Harlow, UK, 1990.

[5] H. Berestycki, B. Larrouturou, and P. L. Lions, *Multidimensional travelling wave solutions of a flame propagation model*, Arch. Rational Mech. Anal., 111 (1990), pp 33–49.

[6] H. Berestycki, B. Larrouturou, J. M. Roquejoffre, *Stability of travelling fronts in a model for flame propagation, part* 1: *linear stability*, Arch. Rational Mech. Anal., 117 (1992), pp. 97–117.

[7] H. Berestycki and L. Nirenberg, *Some qualitative properties of solutions of semilinear equations in cylindrical domains*, in Analysis et Cetera, Rabinowitz and Zehnder, eds., Academic Press, New York, 1990, pp. 115–164.

[8] ———, *Travelling fronts in cylinders*, to appear.

[9] A. Friedman, *Partial differential equations of parabolic type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[10] P. S. Hagan, *Travelling wave and multiple travelling wave solutions of parabolic equations*, SIAM J. Math. Anal, 13 (1992), pp. 717–738.

[11] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lectures Notes in Math., Springer-Verlag, New York, 1981.
[12] K. KIRCHGÄSSNER, *On the nonlinear dynamics of travelling fronts*, J. Differential Equations, to appear.
[13] A. N. KOLMOGOROV, I. G. PETROVSKII, AND N. S. PISKUNOV, *A study of the equation of diffusion with increase in the quantity of matter and its application to a biological problem*, Bjul. Moskovskogo Gos. Univ., 17 (1937), pp. 1–26.
[14] T. LACHAND-ROBERT, to appear.
[15] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[16] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972–1979.
[17] J. M. ROQUEJOFFRE, *Stability of travelling fronts in a model for flame propagation, part 2: nonlinear stability*, Arch. Rational Mech. Anal., 117 (1992), pp. 119–153.
[18] ———, *Convergence to travelling waves for solutions of a class of semilinear parabolic equations*, CMAP internal report 236; J. Differential Equations, to appear.
[19] D. H. SATTINGER, *On the stability of waves of nonlinear parabolic systems*, Adv. Math., 22 (1976), pp. 312–355.
[20] H. B. STEWART, *Generation of analytic semigroups by strongly elliptic operators under general boundary conditions*, Trans. Amer. Math. Soc., 259 (1980), pp. 299–310.
[21] K. UCHIYAMA, *The behavior of solutions of some nonlinear diffusion equations for large time*, J. Math. Kyoto Univ., 18-3 (1978), pp. 453–508.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 21–34, January 1995

002

SLOW MOTION IN ONE-DIMENSIONAL CAHN–MORRAL SYSTEMS*

CHRISTOPHER P. GRANT†

Abstract. In this paper one-dimensional Cahn–Morral systems, which are the multicomponent analogues of the Cahn–Hilliard model for phase separation and coarsening in binary mixtures, are studied. In particular, there is an examination of solutions that start with initial data close to the preferred phases except at finitely many transition points where the data has sharp transition layers, and it is shown that such solutions may evolve exponentially slowly, i.e., if ε is the interaction length then there exists a constant C such that in $\exp(C/\varepsilon)$ units of time the change in such a solution is $o(1)$. This corresponds to extremely slow coarsening of a multicomponent mixture after it has undergone fine-grained decomposition.

Key words. Cahn–Hilliard equation, phase separation, transition layers, metastability

AMS subject classifications. 35B30, 35B25, 35K55

1. Introduction. One of the leading continuum models for the dynamics of phase separation and coarsening in a binary mixture is the Cahn-Hilliard equation, which in the one-dimensional case can be written as

$$\begin{aligned} u_t &= (-\varepsilon^2 u_{xx} + W'(u))_{xx}, && x \in (0,1), \\ u_x &= u_{xxx} = 0, && x \in \{0,1\}. \end{aligned} \tag{1.1}$$

Here W represents the bulk free energy density as a function of the concentration u of one of the two components of the mixture. (If, as is typically assumed, the total concentration of the mixture is a constant then the concentration of the second is determined by the concentration of the first.) The parameter ε represents an interaction length and is assumed to be a small positive constant. This equation was derived in [8] based on the free energy functional of van der Waals [29]

$$\mathcal{E}_\varepsilon[u] \equiv \int_0^1 \left(W(u) + \frac{\varepsilon^2}{2} |u_x|^2 \right) dx. \tag{1.2}$$

We will usually work with the scaled energy $E_\varepsilon[u] \equiv \varepsilon^{-1}\mathcal{E}_\varepsilon[u]$. Also, we will write $E_\varepsilon[u; a, b]$ when the integral is over the interval $[a, b]$ instead of $[0, 1]$.

In the early 1970s, Cahn and Morral [24] and DeFontaine [13], [14] initiated the study of systems of partial differential equations that model the phase separation of mixtures of three or more components in essentially the same way that the Cahn-Hilliard equation models the separation of binary mixtures. (See Eyre [20] for a comprehensive survey of these systems.) If the domain is again taken to be $[0, 1]$, then, after a change of variables, such systems can be written in the form

$$\begin{aligned} u_t &= (-\varepsilon^2 u_{xx} + DW(u))_{xx}, && x \in (0,1), \\ u_x &= u_{xxx} = 0, && x \in \{0,1\}, \end{aligned} \tag{1.3}$$

where u is now an n-vector (for a mixture with $n+1$ components), and W maps $\mathcal{D}(W) \subset \mathbf{R}^n$ into $\mathbf{R}$. Again, $\mathcal{E}_\varepsilon$ defined by (1.2) represents the total free energy of

* Received by the editors February 12, 1992; accepted for publication (in revised form) August 16, 1993.

† Center for Dynamical Systems and Nonlinear Studies, Georgia Institute of Technology, Atlanta, Georgia 30332. Present address, Department of Mathematics, Brigham Young University, Provo, Utah 84602.

the mixture, and it is easy to check that it provides a Lyapunov functional for (1.3). Note, also, that the mass $\int_0^1 u\,dx$ of a solution is conserved.

We will make the following assumptions on W.

- $\mathcal{D}(W)$ is open, convex, and connected;
- $W \geq 0$ throughout its domain, and W has only finitely many zeros, call them $\{z_1, z_2, \ldots, z_m\}$, (corresponding to the preferred homogeneous states, or *phases*, of the system);
- W is C^3 on $\mathcal{D}(W)$ and has a continuous extension to its closure $\overline{\mathcal{D}(W)}$;
- the Hessian D^2W is positive definite at each zero of W, and W is bounded away from 0 outside of each neighborhood of these points.

Additionally, we need to require that W increases as the boundary $\partial\mathcal{D}(W)$ of the domain is approached. The precise assumption we shall make is the following:

- For each point $\overline{u}$ in $\partial\mathcal{D}(W)$, there is a closed, convex set $S \subset \overline{\mathcal{D}(W)} \setminus \overline{u}$ such that we have the following.

1. W is nonzero on Ω, the connected component of $\overline{\mathcal{D}(W)} \setminus S$ containing $\overline{u}$;
2. the function φ that maps each point of $\mathbf{R}^n$ to its nearest point in S satisfies $W(\varphi(u)) \leq W(u)$ for all $u \in \Omega$.

This assumption is trivially satisfied when $\mathcal{D}(W) = \mathbf{R}^n$. It also holds whenever W is C^1 on $\overline{\mathcal{D}(W)}$, $\partial\mathcal{D}(W)$ is a locally compact, oriented hypersurface of class C^2, and the (exterior) normal derivative of W is positive. (See, e.g., [21].) However, we state the assumption in this general way because some of the most important examples of $\mathcal{D}(W)$ do not have smooth boundaries. For example, Eyre [20] and Elliott and Luckhaus [18] study situations where $\mathcal{D}(W)$ is a convex polytope and W satisfies the assumptions given above.

Note that any constant is an equilibrium solution to (1.3). A linear analysis of the equation about an unstable constant equilibrium suggests that typical solutions that start near such a constant undergo fine-grained decomposition with a characteristic length scale that is $O(\varepsilon)$. (See [22] for a precise mathematical formulation and rigorous verification of this heuristic concept in the two-component case.) This fine-grained decomposition of initially homogeneous mixtures has also been frequently observed in physical experiments [7], [9]. In this paper we investigate the way solutions evolve after this initial stage of decomposition. We, therefore, confine our attention to solutions to (1.3) with initial data $u(x,0) = u_0(x)$ close to the zeros of W through most of the domain, with sharp transition layers, or *interfaces*, separating the intervals where u is nearly constant.

Consider when $n = 1$ (i.e., the original Cahn–Hilliard equation (1.1)), the case for which the most work has been done. Carr, Gurtin, and Slemrod [10] showed that all of the local minimizers of E_ε with any specified mass are monotone, so, in general, we would expect that the fine-grained structure of u would coarsen as $t \to \infty$. Numerical work by Elliott and French [17] indicates that this evolution occurs very slowly. (Such slowly evolving states are sometimes said to be *dynamically metastable.*) Bronsard and Hilhorst [5] have shown that, in a certain space, this evolution occurs at a rate that is $O(\varepsilon^k)$ for any power k. Using completely different techniques, Alikakos, Bates, and Fusco [1] constructed a portion of the unstable manifold of a two-layer equilibrium that intersects a small neighborhood of a monotone equilibrium and showed that the speed of the flow along this connecting orbit, measured in the H^{-1} norm, is $O(\exp(-C/\varepsilon))$ for some constant C. Recently, Bates and Xun [4] have found exponentially slow motion for the multilayer states of (1.1) by combining the methods of [1] with those used by Carr and Pego [11] to study reaction-diffusion equations.

The results that we present here are similar to those of Bates and Xun in that we also obtain exponentially slow motion, but the methods we use are much simpler, and they are valid not only for the two-component Cahn–Hilliard equation (1.1) but for the multicomponent Cahn–Morral system (1.3), as well. It should be mentioned, however, that our results for the two-component two-layer case are weaker than those of Alikakos, Bates, and Fusco, in the sense that we do not explicitly construct heteroclinic orbits. We deal only with the speed of motion and say nothing about the geometric structure of the attractor.

In this paper, we apply the elementary yet powerful approach introduced by Bronsard and Kohn [6] in their study of slow motion for reaction-diffusion equations. The improvement from superpolynomial to exponential speed is made possible by incorporating some ideas of Alikakos and McKinney [2] about the profile of constrained minimizers of (1.2). Use is also made of techniques of Sternberg [27] for describing the nature of globally stable steady-state solutions of (1.3) in the limit as $\varepsilon \to 0$.

In §2 we present a lower bound on the energy of any function that is sufficiently close to a given simple function whose range is a subset of $W^{-1}(\{0\})$. This result amounts to an error estimate for a convergence result of Baldo [3]. In §3 we show how this estimate yields our main result on slow evolution of solutions with transition layers. As in [6], the only information used about the time-dependent partial differential equation is the time rate of change of the energy along a solution path in phase space. Finally, in §4 we consider what the main result implies about the motion of the transition layers themselves.

The questions of existence and regularity of solutions for (1.1) and (1.3) have been extensively studied, and different authors have obtained various conditions on W that ensure global existence of solutions [15], [16], [18]–[20], [25], [26], [28], [30]. Rather than restricting ourselves to one particular set of such conditions, we shall simply assume that W is such that for sufficiently smooth initial data with range in $\mathcal{D}(W)$ there exists a global solution that is in $C(\mathbf{R}^+; H^2(0,1)) \cap L^2(0,T; H^4(0,1))$. Given that global solutions exist, our goal is to provide some information about how some of them evolve.

2. Error estimates. Fix $v : [0,1] \to W^{-1}(\{0\})$ having (exactly) N jumps located at $\{x_1, x_2, \dots, x_N\} \subset (0,1)$. Fix r so small that $B(x_k, r) \subset [0,1]$ for each k, and

$$B(x_k, r) \cap B(x_\ell, r) = \emptyset$$

whenever $k \neq \ell$. (Here and below, $B(x,r)$ represents the open ball of radius r centered at x in the relevant space.) Let λ_j be the minimum of the eigenvalues of $D^2W(z_j)$, and let

$$\lambda = \min\{\lambda_j : z_j \in W^{-1}(\{0\})\}.$$

For any function z on $[0,1]$ we write $\tilde{z}(x) \equiv \int_0^x z(s)\,ds$. We are interested in solutions corresponding to initial data $u(x,0) = u_0(x)$ such that $\tilde{u}_0$ is close to $\tilde{v}$ in the L^1 norm. To the discontinuous function v we assign an asymptotic energy

$$E_0[v] \equiv \sum_{k=1}^{N} \phi(v(x_k - r), v(x_k + r)),$$

where

$$\phi(\zeta_1, \zeta_2) \stackrel{\text{def}}{=} \inf\{J[z] : z \in AC([0,1]; \mathcal{D}(W)), z(0) = \zeta_1, z(1) = \zeta_2\},$$

and

$$J[z] \stackrel{\text{def}}{=} \sqrt{2} \int_0^1 \sqrt{W(z(s))}|z'(s)|\, ds.$$

It is easy to check that ϕ is a metric on the domain of W. Also, note that Young's inequality and a change of variable imply that

$$E_\varepsilon[z; a, b] \geq \phi(z(a), z(b)).$$

LEMMA 2.1. *Let C be any positive constant less than $r\sqrt{2\lambda}$. Then there are constants $C_1, \delta > 0$ (depending only on W, v and C) such that, for ε sufficiently small,*

$$\int_0^1 |\tilde{u}(x) - \tilde{v}(x)|\, dx \leq \delta \quad \Rightarrow \quad E_\varepsilon[u] \geq E_0[v] - C_1 \exp(-C/\varepsilon).$$

Proof. Let K be a compact set in the domain of W containing $W^{-1}(\{0\})$ in its interior, and set $\kappa = \sup\{\|D^3W(\zeta)\| : \zeta \in K\}$. Choose $\hat{r} > 0$ and ρ_1 so small that $C \leq (r - \hat{r})\sqrt{2\lambda - n\kappa\rho_1}$ and that $B(z_j, \rho_1)$ is contained in K for each $z_j \in W^{-1}(\{0\})$. Choose ρ_2 so small that

$$\begin{aligned} &\inf\left\{\phi(\zeta_1, \zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_1 \notin B(z_j, \rho_1), \zeta_2 \in B(z_j, \rho_2)\right\} \\ &> \sup\left\{\phi(z_j, \zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_2 \in B(z_j, \rho_2)\right\}, \end{aligned}$$

and $|z_j - z_\ell| > 2\rho_2$ if z_j and z_ℓ are different zeros of W.

Let

$$\begin{aligned} F(\rho_2) = \inf\{&\phi(\zeta_1, \zeta_2) : z_{j_1}, z_{j_2} \in W^{-1}(\{0\}), z_{j_1} \neq z_{j_2}, \\ &\zeta_1 \in B(z_{j_1}, \rho_2), |(\zeta_2 - z_{j_2}) \cdot (z_{j_2} - z_{j_1})| \leq \rho_2 |z_{j_2} - z_{j_1}|\}. \end{aligned} \tag{2.1}$$

By our assumptions about W, $F(\rho_2) > 0$, so there exists $M \in \mathbf{N}$ such that $MF(\rho_2) > E_0[v]$. Pick such an M, and set $\delta = \hat{r}^2\rho_2/(5M^2)$.

Now assume that $\int_0^1 |\tilde{u}(x) - \tilde{v}(x)|\, dx \leq \delta$, and let us focus our attention on $B(x_k, r)$, a neighborhood of one of the transition points of v. For convenience, let $v_+ = v(x_k + r)$ and $v_- = v(x_k - r)$. Suppose $|u - v| \geq \rho_2$ throughout $(x_k, x_k + \hat{r})$, and let I_M be an open subinterval of $(x_k, x_k + \hat{r})$ of width $\hat{r}/M$. If we assume without loss of generality that $E_\varepsilon[u] \leq E_0[v]$ then for ε sufficiently small there must be some $\hat{x} \in I_M$ such that $u(\hat{x}) \in B(z_{j_1}, \rho_2)$ for some $z_{j_1} \in W^{-1}(\{0\})$. (Otherwise the rescaled bulk free energy would be too high.) If

$$\left|(u - v) \cdot \frac{z_{j_1} - v_+}{|z_{j_1} - v_+|}\right| \geq \rho_2$$

throughout I_M then it is not hard to check that we would have

$$\int_{I_M} |\tilde{u}(x) - \tilde{v}(x)|\, dx \geq \int_{I_M} \left|(\tilde{u}(x) - \tilde{v}(x)) \cdot \frac{z_{j_1} - v_+}{|z_{j_1} - v_+|}\right| dx > \delta,$$

which is a contradiction. Hence,

$$\left|(u - v) \cdot \frac{z_{j_1} - v_+}{|z_{j_1} - v_+|}\right| < \rho_2$$

somewhere on I_M. But then the rescaled energy on I_M must be no less than $F(\rho_2)$. Partitioning $(x_k, x_k + \hat{r})$ into M equal intervals of width $\hat{r}/M$ and using the preceding result, we have $E_\varepsilon[u; x_k, x_k + \hat{r}] \geq MF(\rho_2) > E_0[v]$, contrary to assumption. Hence, there is some $r_+ \in (0, \hat{r})$ such that

$$|u(x_k + r_+) - v_+| < \rho_2.$$

Similarly, there is some $r_- \in (0, \hat{r})$ such that

$$|u(x_k - r_-) - v_-| < \rho_2.$$

Next, consider the unique minimizer $z : [x_k + r_+, x_k + r] \to \mathbf{R}^n$ of the functional $E_\varepsilon[z; x_k + r_+, x_k + r]$ subject to the boundary condition

$$z(x_k + r_+) = u(x_k + r_+).$$

If the range of z is not contained in $B(v_+, \rho_1)$ then

$$\begin{aligned} E_\varepsilon[z; x_k + r_+, x_k + r] &\geq \inf\{\phi(z(x_k + r_+), \zeta) : \zeta \notin B(v_+, \rho_1)\} \\ &\geq \phi(z(x_k + r_+), v_+), \end{aligned} \tag{2.2}$$

by the choice of ρ_2 and the choice of r_+.

Suppose, on the other hand, that the range of z is contained in $B(v_+, \rho_1)$. Then the Euler–Lagrange equation for z is

$$\begin{aligned} z''(x) &= \varepsilon^{-2} DW(z(x)), & x &\in (x_k + r_+, x_k + r), \\ z(x) &= u(x_k + r_+), & x &= x_k + r_+, \\ z'(x) &= 0, & x &= x_k + r. \end{aligned} \tag{2.3}$$

If we define $\psi(x) \equiv |z(x) - v_+|^2$ then $\psi' = 2(z - v_+) \cdot z'$ and

$$\psi'' = 2(|z'|^2 + (z - v_+) \cdot z'') \geq \frac{2}{\varepsilon^2}(z - v_+) \cdot DW(z). \tag{2.4}$$

Now Taylor's theorem and the choice of ρ_1 imply that

$$DW(z) = D^2W(v_+)(z - v_+) + R, \tag{2.5}$$

where $|R| \leq n\kappa|z - v_+|^2/2$. Substituting (2.5) into (2.4) gives

$$\begin{aligned} \psi'' &\geq \frac{2}{\varepsilon^2}(z - v_+) \cdot D^2W(v_+)(z - v_+) - \frac{n\kappa}{\varepsilon^2}|z - v_+|^3 \\ &\geq \frac{2\lambda}{\varepsilon^2}|z - v_+|^2 - \frac{n\kappa\rho_1}{\varepsilon^2}|z - v_+|^2 \\ &\geq \frac{\mu^2}{\varepsilon^2}|z - v_+|^2 \\ &= \frac{\mu^2}{\varepsilon^2}\psi, \end{aligned}$$

where $\mu = C/(r - \hat{r})$.

Thus, ψ satisfies

$$\begin{aligned} \psi''(x) - (\mu/\varepsilon)^2\psi(x) &\geq 0, & x &\in (x_k + r_+, x_k + r), \\ \psi(x) &= |u(x_k + r_+) - v_+|^2, & x &= x_k + r_+, \\ \psi'(x) &= 0, & x &= x_k + r. \end{aligned}$$

Following Alikakos and McKinney [2], we compare ψ to the solution $\hat{\psi}$ of

$$\begin{aligned} \hat{\psi}''(x) - (\mu/\varepsilon)^2 \hat{\psi}(x) &= 0, & x &\in (x_k + r_+, x_k + r), \\ \hat{\psi}(x) &= |u(x_k + r_+) - v_+|^2, & x &= x_k + r_+, \\ \hat{\psi}'(x) &= 0, & x &= x_k + r, \end{aligned}$$

which can be explicitly calculated to be

$$\hat{\psi}(x) = \frac{|u(x_k + r_+) - v_+|^2}{\cosh\left[(\mu/\varepsilon)(r - r_+)\right]} \cosh\left[\frac{\mu}{\varepsilon}(x - (x_k + r))\right].$$

By the maximum principle, $\psi(x) \le \hat{\psi}(x)$, so, in particular,

$$\psi(x_k + r) \le \frac{|u(x_k + r_+) - v_+|^2}{\cosh\left[(\mu/\varepsilon)(r - r_+)\right]} \le 2\rho_2^2 \exp\left[-\frac{C}{\varepsilon}\right].$$

Consequently,

$$|z(x_k + r) - v_+| \le \rho_2 \sqrt{2} \exp(-C/(2\varepsilon)). \tag{2.6}$$

Because W is quadratic at v_+, (2.6) implies that, for some constant C_1,

$$\begin{aligned} E_\varepsilon[z; x_k + r_+, x_k + r] &\ge \phi(z(x_k + r_+), z(x_k + r)) \\ &\ge \phi(z(x_k + r_+), v_+) - \phi(v_+, z(x_k + r)) \\ &\ge \phi(z(x_k + r_+), v_+) - (C_1/(2N)) \exp(-C/\varepsilon). \end{aligned} \tag{2.7}$$

Combining (2.2) and (2.7), we see that the constrained minimizer of the proposed variational problem satisfies

$$E_\varepsilon[z; x_k + r_+, x_k + r] \ge \phi(z(x_k + r_+), v_+) - (C_1/(2N)) \exp(-C/\varepsilon).$$

But the restriction of u to $[x_k + r_+, x_k + r]$ is an admissable function, so it must satisfy the same estimate

$$E_\varepsilon[u; x_k + r_+, x_k + r] \ge \phi(u(x_k + r_+), v_+) - (C_1/(2N)) \exp(-C/\varepsilon).$$

A similar estimate holds for the energy of u on the interval $[x_k - r, x_k - r_-]$. Hence,

$$\begin{aligned} E_\varepsilon[u; x_k - r, x_k + r] &= E_\varepsilon[u; x_k - r, x_k - r_-] + E_\varepsilon[u; x_k - r_-, x_k + r_+] \\ &\quad + E_\varepsilon[u; x_k + r_+, x_k + r] \\ &\ge \phi(v_-, u(x_k - r_-)) - (C_1/(2N)) \exp(-C/\varepsilon) \\ &\quad + \phi(u(x_k - r_-), u(x_k + r_+)) \\ &\quad + \phi(u(x_k + r_+), v_+) - (C_1/(2N)) \exp(-C/\varepsilon) \\ &\ge \phi(v(x_k - r), v(x_k + r)) - (C_1/N) \exp(-C/\varepsilon). \end{aligned}$$

Assembling all of our estimates,

$$E_\varepsilon[u] \ge \sum_{k=1}^{N} E_\varepsilon[u; x_k - r, x_k + r] \ge E_0[v] - C_1 \exp(-C/\varepsilon),$$

as was claimed. □

3. Slow evolution. In this section we will consider a family of solutions $u^\varepsilon(x,t)$ to (1.3), parametrized by the corresponding interaction length ε.

LEMMA 3.1. *Suppose that $C < r\sqrt{2\lambda}$ and the initial data u_0^ε satisfies*

$$\int_0^1 |\tilde{u}_0^\varepsilon(x) - \tilde{v}(x)|\,dx \le \frac{\delta}{2}$$

and

$$E_\varepsilon[u_0^\varepsilon] \le E_0[v] + \frac{1}{g(\varepsilon)}$$

for some function g and for all ε small, where δ is as in Lemma 2.1. Then

$$\lim_{\varepsilon\to 0}\left\{\sup_{0\le t\le \min\{g(\varepsilon),\exp(C/\varepsilon)\}}\int_0^1 |\tilde{u}^\varepsilon(x,t) - \tilde{u}_0^\varepsilon(x)|\,dx\right\} = 0. \tag{3.1}$$

Proof. First note that the scaled total energy $E_\varepsilon[u^\varepsilon(\cdot,t)]$ of the solution of a Cahn-Morral system is nonincreasing in t, since

$$\begin{aligned}\frac{d}{dt}E_\varepsilon[u^\varepsilon(\cdot,t)] &= \varepsilon^{-1}\int_0^1 \left[DW(u^\varepsilon)\cdot u_t^\varepsilon + \varepsilon^2 u_x^\varepsilon \cdot u_{xt}^\varepsilon\right] dx\\ &= \varepsilon^{-1}\int_0^1 \left[(DW(u^\varepsilon) - \varepsilon^2 u_{xx}^\varepsilon)\cdot u_t^\varepsilon\right] dx\\ &= -\varepsilon^{-1}\int_0^1 |\tilde{u}_t^\varepsilon|^2\,dx.\end{aligned}$$

Integrating this equation over $t \in (0,T)$ gives

$$E_\varepsilon[u_0^\varepsilon] - E_\varepsilon[u^\varepsilon(\cdot,T)] = \varepsilon^{-1}\int_0^T\int_0^1 |\tilde{u}_t^\varepsilon|^2\,dx\,dt. \tag{3.2}$$

Next, assume that u_0^ε satisfies the conditions of the lemma and that T is small enough that

$$\int_0^T\int_0^1 |\tilde{u}_t^\varepsilon|\,dx\,dt \le \delta/2.$$

Then

$$\int_0^1 |\tilde{u}_0^\varepsilon(x) - \tilde{u}^\varepsilon(x,T)|\,dx \le \delta/2,$$

so by the triangle inequality,

$$\int_0^1 |\tilde{u}^\varepsilon(x,T) - \tilde{v}(x)|\,dx \le \delta.$$

Applying, Lemma 2.1 to $\tilde{u}^\varepsilon(\cdot,T)$ gives $E_\varepsilon[u^\varepsilon(\cdot,T)] \ge E_0[v] - C_1\exp(-C/\varepsilon)$. In combination with (3.2), this yields

$$\begin{aligned}\int_0^T\int_0^1 |\tilde{u}_t^\varepsilon|^2\,dx\,dt &= \varepsilon(E_\varepsilon[u_0^\varepsilon] - E_\varepsilon[u^\varepsilon(\cdot,T)])\\ &\le C_1\varepsilon\left[\frac{1}{g(\varepsilon)} + \exp(-C/\varepsilon)\right],\end{aligned} \tag{3.3}$$

assuming, without loss of generality, that $C_1 \geq 1$.

Using Hölder's inequality and (3.3) we have

$$\left(\int_0^T \int_0^1 |\tilde{u}_t^\varepsilon| \, dx \, dt\right)^2 \leq \left(\int_0^T \int_0^1 1 \, dx \, dt\right) \cdot \left(\int_0^T \int_0^1 |\tilde{u}_t^\varepsilon|^2 \, dx \, dt\right)$$
$$\leq C_1 T \varepsilon \left[\frac{1}{g(\varepsilon)} + \exp(-C/\varepsilon)\right].$$

Hence,

$$T \geq \frac{1}{C_1 \varepsilon} \left[\frac{1}{g(\varepsilon)} + \exp(-C/\varepsilon)\right]^{-1} \left(\int_0^T \int_0^1 |\tilde{u}_t^\varepsilon| \, dx \, dt\right)^2. \tag{3.4}$$

Now suppos that

$$\int_0^\infty \int_0^1 |\tilde{u}_t^\varepsilon| \, dx \, dt \geq \delta/2.$$

Then we can choose T such that $\int_0^T \int_0^1 |\tilde{u}_t^\varepsilon| \, dx \, dt = \delta/2$. For this choice of T, equation (3.4) yields

$$T \geq \frac{\delta^2}{4C_1 \varepsilon \left[\frac{1}{g(\varepsilon)} + \exp(-C/\varepsilon)\right]} \geq \frac{\delta^2}{8C_1 \varepsilon} \min\left\{g(\varepsilon), \exp(C/\varepsilon)\right\}.$$

Then (3.3) implies that

$$\int_0^{\delta^2 \min\{g(\varepsilon), \exp(C/\varepsilon)\}/(8C_1\varepsilon)} \int_0^1 |\tilde{u}_t^\varepsilon|^2 \, dx \, dt \leq C_1 \varepsilon \left[\frac{1}{g(\varepsilon)} + \exp(-C/\varepsilon)\right]. \tag{3.5}$$

If, on the other hand, $\int_0^\infty \int_0^1 |\tilde{u}_t^\varepsilon| \, dx \, dt < \delta/2$, then (3.3) must hold for every T; therefore, (3.5) is also true for this case.

Using Hölder's inequality and (3.5) we see that for $\varepsilon < \delta^2/(8C_1)$

$$\begin{aligned}
&\sup_{0 \leq t \leq \min\{g(\varepsilon), \exp(C/\varepsilon)\}} \int_0^1 |\tilde{u}^\varepsilon(x,t) - \tilde{u}_0^\varepsilon(x)| \, dx \\
&\leq \int_0^{\min\{g(\varepsilon), \exp(C/\varepsilon)\}} \int_0^1 |\tilde{u}_t^\varepsilon| \, dx \, dt \\
&\leq \left(\min\left\{g(\varepsilon), \exp(C/\varepsilon)\right\} \int_0^{\min\{g(\varepsilon), \exp(C/\varepsilon)\}} \int_0^1 |\tilde{u}_t^\varepsilon|^2 \, dx \, dt\right)^{1/2} \\
&\leq \left(\min\left\{g(\varepsilon), \exp(C/\varepsilon)\right\} C_1 \varepsilon \left[\frac{1}{g(\varepsilon)} + \exp(-C/\varepsilon)\right]\right)^{1/2} \\
&\leq \sqrt{2C_1 \varepsilon}.
\end{aligned}$$

Letting $\varepsilon \to 0$ we get (3.1). □

The strength of estimate (3.1) in Lemma 3.1 depends on the efficiency of the transition layers in the initial data. In Theorem 3.3 below, we show that, in a neighborhood of the step function v, there exist initial data that smooth out the discontinuities of

v in an efficient enough manner that the corresponding solutions of (1.3) evolve in an exponentially slow way. Before we present this theorem, we shall state and prove a technical lemma about the existence and regularity of minimizing geodesics for the degenerate Riemannian metric ϕ.

LEMMA 3.2. (1) *For any two zeros z_i and z_j of W, there is a Lipschitz continuous path γ_{ij} from z_i to z_j, parametrized by a multiple of Euclidean arclength, that realizes the distance $\phi(z_i, z_j)$*; i.e., $\phi(z_i, z_j) = J[\gamma_{ij}]$.

(2) *There exists a positive constant C_2 such that $|\gamma_{ij}(y) - z_i| \geq C_2 y$ for y sufficiently small, and $|\gamma_{ij}(y) - z_j| \geq C_2(1-y)$ for y sufficiently near* 1.

Proof. Recall that outside of a neighborhood of its zeros W is bounded away from 0; therefore, it is possible to find a bounded set $B \subset \mathcal{D}(W)$ such that if $\gamma(0) = z_i$, $\gamma(1) = z_j$, and $J[\gamma] \leq \phi(z_i, z_j) + 1$ then the image of γ is contained in B. Extend W continuously to $\overline{B}$, and consider the problem of minimizing $J[\gamma]$ over all γ satisfying these boundary conditions and having images contained in $\overline{B}$. Now, $J[\gamma]$ is a parametric integral, and it is known that this new minimization problem has an AC global minimizer γ_{ij} [12]. The parameter of this minimizer can be chosen to be proportional to arclength, and then γ_{ij} will be Lipschitz continuous.

We claim that $\gamma_{ij}([0,1])$ is contained in $\mathcal{D}(W)$. Suppose it is not. Then there exists some $y \in (0,1)$ such that $\gamma_{ij}(y) \in \partial\mathcal{D}(W)$. By the assumptions on W, there is a closed, convex set $S \subset \overline{\mathcal{D}(W)} \setminus \gamma_{ij}(y)$ such that W is nonzero on the connected component Ω of $\overline{\mathcal{D}(W)} \setminus S$ containing $\gamma_{ij}(y)$, and the function φ that maps each point of $\mathbf{R}^n$ to its nearest point in S satisfies $W(\varphi(u)) \leq W(u)$ for all u in Ω. Consider the modified path $\bar{\gamma}_{ij}$ from z_i to z_j defined by

$$\bar{\gamma}_{ij}(y) = \begin{cases} \varphi(\gamma_{ij}(y)) & \text{if } \gamma_{ij}(y) \in \Omega, \\ \gamma_{ij}(y) & \text{otherwise.} \end{cases}$$

Note that φ is Lipschitz continuous with Lipschitz constant 1. Because of this and the fact that S separates Ω from the rest of $\overline{\mathcal{D}(W)}$, $\bar{\gamma}_{ij}$ is Lipschitz continuous. It is also easy to check that $J[\bar{\gamma}_{ij}] < J[\gamma_{ij}]$. This contradicts the optimality of γ_{ij}; hence, the claim holds. This verifies that $\phi(z_i, z_j) = J[\gamma_{ij}]$.

We now prove the estimate on γ_{ij} near z_i; the estimate near z_j can be derived similarly. Again, we consider a modification of γ_{ij}, this time the path γ_{ij}^{η} defined by

$$\gamma_{ij}^{\eta}(y) = \begin{cases} z_i + (y/\eta)(\gamma_{ij}(\eta) - z_i) & \text{if } 0 \leq y \leq \eta, \\ \gamma_{ij}(y) & \text{otherwise.} \end{cases}$$

The optimality of γ_{ij} implies that

$$\sqrt{2}\int_0^{\eta} \sqrt{W(\gamma_{ij}(s))}|\gamma_{ij}'(s)|\,ds \leq \sqrt{2}\int_0^{\eta} \sqrt{W(\gamma_{ij}^{\eta}(s))}\frac{|\gamma_{ij}(\eta) - z_i|}{\eta}\,ds. \tag{3.6}$$

Because $D^2W(z_i)$ is positive definite, there are positive constants M_1 and M_2 such that

$$M_1|u - z_i| \leq \sqrt{W(u)} \leq M_2|u - z_i|$$

in a small neighborhood of z_i. Using this in (3.6), we find that

$$|\gamma_{ij}(\eta) - z_i|^2 \geq M_3 \int_0^{\eta} |\gamma_{ij}(s) - z_i|\,ds$$

for some constant M_3. Applying a variant of Gronwall's inequality [23] we obtain the desired estimate. □

THEOREM 3.3. *Given $\delta > 0$, there exist constants $C, \hat{\varepsilon} > 0$ and a family of initial conditions $\{u_0^\varepsilon : 0 \le \varepsilon \le \hat{\varepsilon}\}$ of (1.3) satisfying homogeneous Neumann boundary conditions and the estimate*

$$\int_0^1 |\tilde{u}_0^\varepsilon(x) - \tilde{v}(x)|\, dx \le \frac{\delta}{2}$$

such that the corresponding solutions u^ε of (1.3) satisfy

$$\lim_{\varepsilon \to 0} \left\{ \sup_{0 \le t \le \exp(C/\varepsilon)} \int_0^1 |\tilde{u}^\varepsilon(x,t) - \tilde{u}_0^\varepsilon(x)|\, dx \right\} = 0.$$

Proof. Lemma 3.2 shows that to each discontinuity x_k of v there corresponds an optimal path connecting $v(x_k - r)$ to $v(x_k + r)$. Note that it suffices to prove the present theorem under the assumption that none of these optimal paths passes through any zero of W (except at the endpoints of the path), since if the assumption is not satisfied then v can be perturbed slightly to create a new step function that does satisfy the assumption.

Given ε, set $u_0^\varepsilon = v$ outside of $\cup_{j=1}^m B(x_k, r)$. For fixed x_k, we shall again use the notation $v_\pm$ for $v(x_k \pm r)$ and will show that for ε sufficiently small we can define u_0^ε inside $B(x_k, r)$ in such a way that u_0^ε is very close to v (in the L^1 sense) on $B(x_k, r)$, $E_\varepsilon[u_0^\varepsilon; x_k - r, x_k + r] \le \phi(v_-, v_+) + C_3 \exp(-C/\varepsilon)$ for some C and C_3, and u_0^ε is continuous at the endpoints of $B(x_k, r)$. By taking C slightly smaller and applying Lemma 3.1, the proof of the theorem will then be complete.

Let $\gamma : [0,1] \to \mathbf{R}^n$ be an optimal path from v_- to v_+ as described in Lemma 3.2. Let σ be the Euclidean arclength of γ. Let $y : \mathbf{R} \to [0,1]$ be the solution of

$$\frac{dy}{d\xi} = \sigma^{-1} \sqrt{2W(\gamma(y(\xi)))} \tag{3.7}$$

satisfying $y(0) = 1/2$. (Since $\sqrt{W}$ and γ are Lipschitz continuous, a unique C^1 solution is guaranteed to exist.) Note that $\lim_{\xi \to \infty} y(\xi) = 1$ and $\lim_{\xi \to -\infty} y(\xi) = 0$. Define u_0^ε inside $B(x_k, r)$ by

$$u_0^\varepsilon(x) = \begin{cases} v_- + (\gamma(y(1 - r/\varepsilon)) - v_-)(x - x_k + r)/\varepsilon, & x_k - r < x < x_k - r + \varepsilon, \\ \gamma(y((x - x_k)/\varepsilon)), & x_k - r + \varepsilon \le x \le x_k + r - \varepsilon, \\ v_+ + (v_+ - \gamma(y(r/\varepsilon - 1)))(x - x_k - r)/\varepsilon, & x_k + r - \varepsilon < x < x_k + r. \end{cases}$$

It is easy to see that u_0^ε is continuous and, for ε sufficiently small, will satisfy the L^1 requirement; therefore, we only need to check the energy requirement. Note that

$$\begin{aligned} E_\varepsilon[u_0^\varepsilon; x_k - r, x_k + r] &= \int_{-r}^{-r+\varepsilon} \left[\frac{1}{\varepsilon} W(u_0^\varepsilon(x + x_k)) + \frac{\varepsilon}{2} |u_0^{\varepsilon\prime}(x + x_k)|^2 \right] dx \\ &\quad + \int_{-r+\varepsilon}^{r-\varepsilon} \left[\frac{1}{\varepsilon} W(u_0^\varepsilon(x + x_k)) + \frac{\varepsilon}{2} |u_0^{\varepsilon\prime}(x + x_k)|^2 \right] dx \\ &\quad + \int_{r-\varepsilon}^{r} \left[\frac{1}{\varepsilon} W(u_0^\varepsilon(x + x_k)) + \frac{\varepsilon}{2} |u_0^{\varepsilon\prime}(x + x_k)|^2 \right] dx \\ &\stackrel{\text{def}}{=} I_1 + I_2 + I_3. \end{aligned} \tag{3.8}$$

Now, using (3.7) and the definition of γ we have

$$
\begin{aligned}
I_2 &= \int_{-r+\varepsilon}^{r-\varepsilon} \left[\frac{1}{\varepsilon} W\left(\gamma\left(y\left(\frac{x}{\varepsilon}\right)\right)\right) + \frac{1}{2\varepsilon}\left|\gamma'\left(y\left(\frac{x}{\varepsilon}\right)\right) y'\left(\frac{x}{\varepsilon}\right)\right|^2 \right] dx \\
&= \int_{1-r/\varepsilon}^{r/\varepsilon-1} \left[W(\gamma(y(\xi))) + \frac{1}{2}|\gamma'(y(\xi))y'(\xi)|^2 \right] d\xi \\
&= \int_{1-r/\varepsilon}^{r/\varepsilon-1} \sqrt{2W(\gamma(y(\xi)))}|\gamma'(y(\xi))|y'(\xi)\, d\xi \qquad (3.9) \\
&= \int_{y(1-r/\varepsilon)}^{y(r/\varepsilon-1)} \sqrt{2W(\gamma(y))}|\gamma'(y)|\, dy \\
&\le \phi(v_-, v_+).
\end{aligned}
$$

Next, we estimate I_1 (letting C_3 represent a constant whose value may change from line to line):

$$
\begin{aligned}
I_1 &= \frac{1}{\varepsilon}\int_{-r}^{-r+\varepsilon} W\left(v_- + \frac{\gamma\left(y\left(1-r/\varepsilon\right)\right) - v_-}{\varepsilon}(x+r)\right) dx \\
&\quad + \frac{1}{2}\left|\gamma\left(y\left(1-\frac{r}{\varepsilon}\right)\right) - v_-\right|^2 \qquad (3.10) \\
&\le C_3\left|\gamma\left(y\left(1-\frac{r}{\varepsilon}\right)\right) - v_-\right|^2 \\
&\le C_3\left(y\left(1-\frac{r}{\varepsilon}\right)\right)^2.
\end{aligned}
$$

Now, Lemma 3.2 implies that there exists a constant $C > 0$ such that, for $\xi \ll 0$,

$$
\begin{aligned}
y'(\xi) &= \sigma^{-1}\sqrt{2W(\gamma(y(\xi)))} \\
&\ge \frac{C}{2rC_2}|\gamma(y(\xi)) - v_-| \qquad (3.11) \\
&\ge \frac{C}{2r}(y(\xi)).
\end{aligned}
$$

Applying a simple comparison argument to (3.11) yields

$$
y(\xi) \le C_3 \exp\left(\frac{C\xi}{2r}\right)
$$

for $\xi \ll 0$. Substituting this into (3.10) we have

$$
I_1 \le C_3 \exp(-C/\varepsilon). \qquad (3.12)
$$

Similarly,

$$
I_3 \le C_3 \exp(-C/\varepsilon). \qquad (3.13)
$$

By substituting (3.9), (3.12), and (3.13) into (3.8), we see that u_0^ε satisfies the energy requirement, so we are done. □

Remark. For the standard two-component case with W having two minima, the maximum principle can be used more directly in the proof of Lemma 2.1 (see [2])

and an explicit value of C can be obtained in Theorem 3.3. This C agrees with that obtained in [1] and [4].

Remark. The initial data u_0^ε just constructed are in $W^{1,\infty}(0,1)$. Since E_ε is continuous on this space and elements of this space can be approximated arbitrarily closely by C^p functions (for arbitrarily large p), the initial data in Theorem 3.3 can be assumed to be arbitrarily smooth.

4. Motion of transition layers. From Theorem 3.3, which establishes slow evolution in a certain abstract space, it is natural to infer that the movement of the transition layers themselves is extremely slow. This concept can be made precise in a number of ways, one of which we present here.

Fix some closed subset K of $\mathcal{D}(W) \setminus W^{-1}(\{0\})$, and define the *interface* $I[u]$ of a function u by

$$I[u] \stackrel{\text{def}}{=} u^{-1}(K).$$

This terminology is natural, since the set K is bounded away from the phases of W, where the bulk energy is low. By analyzing how rapidly $I[u]$ changes, we obtain information on how fast the transition layers move.

Let $d(A,B)$ denote the Hausdorff distance between the sets A and B, i.e.,

$$d(A,B) = \max\left\{\sup_{a\in A} d(a,B), \sup_{b\in B} d(b,A)\right\}.$$

We shall show that $d(I[u^\varepsilon(\cdot,t)], I[u_0^\varepsilon])$ grows very slowly in t.

THEOREM 4.1. *Fix $\hat{r} > 0$ and $\hat{\delta} > 0$. Then there exist constants $C, \hat{\varepsilon} > 0$ and a family of initial conditions $\{u_0^\varepsilon : 0 \le \varepsilon \le \hat{\varepsilon}\}$ of (1.3) satisfying homogeneous Neumann boundary conditions and the estimate*

$$\int_0^1 |\tilde{u}_0^\varepsilon(x) - \tilde{v}(x)|\, dx \le \hat{\delta}$$

such that the time $T(\hat{r})$ necessary for $d(I[u^\varepsilon(\cdot,T(\hat{r}))], I[u_0^\varepsilon])$ to exceed $\hat{r}$ satisfies

$$T(\hat{r}) \ge \exp(C/\varepsilon). \tag{4.1}$$

Proof. Assume, without loss of generality, that $\hat{r} \le r$. Choose $\hat{\rho}$ small enough that

$$\begin{aligned}&\inf\left\{\phi(\zeta_1,\zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_1 \in K, \zeta_2 \in B(z_j,\hat{\rho})\right\}\\ &\quad > 4N \sup\left\{\phi(z_j,\zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_2 \in B(z_j,\hat{\rho})\right\}.\end{aligned}$$

Choose $M \in \mathbf{N}$ so large that $MF(\hat{\rho}) > E_0[v]$, where F is defined as in (2.1).

We claim that there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \le \varepsilon_0$ and for all functions $z : [0,1] \to \mathbf{R}^n$ satisfying

$$\int_0^1 |\tilde{z}(x) - \tilde{v}(x)|\, dx \le \frac{\hat{\rho}\hat{r}^2}{17M^2} \tag{4.2}$$

and

$$E_\varepsilon[z] \le E_0[v] + 2N \sup\left\{\phi(z_j,\zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_2 \in B(z_j,\hat{\rho})\right\} \tag{4.3}$$

we have

$$d\left(I[z], \{x_k\}_{k=1}^N\right) < \frac{\hat{r}}{2}.$$

Verification of claim. Note first that if ε is sufficiently small then for each k there exist

$$x_{k-} \in (x_k - \hat{r}/2, x_k)$$

and

$$x_{k+} \in (x_{k+}, x_k + \hat{r}/2)$$

such that $|z(x_{k\pm}) - v(x_{k\pm})| < \hat{\rho}$. This follows as in the proof of Lemma 2.1. Now, suppose the claim is violated. Then, reasoning as before,

$$\begin{aligned} E_\varepsilon[z] &\geq \sum_{k=1}^N E_\varepsilon[z; x_{k-}, x_{k+}] \\ &\quad + \inf\left\{\phi(\zeta_1, \zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_1 \in K, \zeta_2 \in B(z_j, \hat{\rho})\right\} \\ &\geq E_0[v] - 2N \sup\left\{\phi(z_j, \zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_2 \in B(z_j, \hat{\rho})\right\} \\ &\quad + \inf\left\{\phi(\zeta_1, \zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_1 \notin K, \zeta_2 \in B(z_j, \hat{\rho})\right\} \\ &> E_0[v] + 2N \sup\left\{\phi(z_j, \zeta_2) : z_j \in W^{-1}(\{0\}), \zeta_2 \in B(z_j, \hat{\rho})\right\} \\ &\geq E_\varepsilon[z], \end{aligned}$$

which is a contradiction. Thus, the claim is true.

Apply Theorem 3.3 with $\delta = \min\{\hat{\delta}, \hat{\rho}\hat{r}^2/(17M^2)$ to obtain a parametrized set of initial conditions $\{u_0^\varepsilon : 0 \leq \varepsilon \leq \hat{\varepsilon}\}$. Note that $z = u_0^\varepsilon$ satisfies (4.2) and, by the construction in the proof of Theorem 3.3, satisfies (4.3) for ε sufficiently small. Applying the claim we get

$$d\left(I[u_0^\varepsilon], \{x_k\}_{k=1}^N\right) < \frac{\hat{r}}{2},$$

for ε sufficiently small. By Theorem 3.3, the triangle inequality, and the fact that $E_\varepsilon[u^\varepsilon(\cdot, t)]$ is decreasing in t, we see that there is a constant $C > 0$ such that for ε sufficiently small, $z = u^\varepsilon(\cdot, T)$ satisfies (4.2) and (4.3) if $T \leq \exp(C/\varepsilon)$. Thus, for all such T we also have

$$d\left(I[u^\varepsilon(\cdot, T)], \{x_k\}_{k=1}^N\right) < \frac{\hat{r}}{2}.$$

By the triangle inequality we get

$$d\left(I[u_0^\varepsilon], I[u^\varepsilon(\cdot, T)]\right) < \hat{r}.$$

This means that (4.1) must hold. □

Acknowledgments. I am indebted to K. Mischaikow for suggesting this problem and to P. Bates for helping to correct a mistake in the original manuscript. I would also like to thank the referees for their helpful suggestions.

REFERENCES

[1] N. D. ALIKAKOS, P. W. BATES, AND G. FUSCO, *Slow motion for the Cahn-Hilliard equation in one space dimension*, J. Differential Equations, 90 (1991), pp. 81–135.

[2] N. D. ALIKAKOS AND W. R. MCKINNEY, *Remarks on the equilibrium theory for the Cahn-Hilliard equation in one space dimension*, in Reaction-Diffusion Equations, K. J. Brown and A. A. Lacey, ed., Oxford University Press, London, 1990.

[3] S. BALDO, *Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluids*, Ann. Inst. H. Poincaré Anal. Nonlinèaire, 7 (1990), pp. 67–90.

[4] P. W. BATES AND J. P. XUN, *Metastable patterns for the Cahn-Hilliard equation*, J. Differential Equations, 111 (1994), pp. 421–457.

[5] L. BRONSARD AND D. HILHORST, *On the slow dynamics for the Cahn-Hilliard equation in one space dimension*, Proc. Roy. Soc. London Ser. A, 439 (1992), pp. 669–682.

[6] L. BRONSARD AND R. V. KOHN, *On the slowness of phase boundary motion in one space dimension*, Comm. Pure Appl. Math., 43 (1990), pp. 983–998.

[7] E. P. BUTLER AND G. THOMAS, *Structure and properties of spinodally decomposed Cu-Ni-Fe alloys*, Acta Metall., 18 (1970), pp. 347–365.

[8] J. W. CAHN, *On spinodal decomposition*, Acta Metall., 9 (1961), pp. 795–801.

[9] ———, *Spinodal decomposition*, Trans. Metallurg. Soc. of AIME, 242 (1968), pp. 166–180.

[10] J. CARR, M. E. GURTIN, AND M. SLEMROD, *Structured phase transitions on a finite interval*, Arch. Rational Mech. Anal., 86 (1984), pp. 317–351.

[11] J. CARR AND R. L. PEGO, *Metastable patterns in solutions of* $u_t = \varepsilon^2 u_{xx} - f(u)$, Comm. Pure Appl. Math., 42 (1989), pp. 523–576.

[12] L. CESARI, *Optimization—Theory and Applications*, Appl. Math., Vol. 17, Springer-Verlag, New York, 1983.

[13] D. DEFONTAINE, *An analysis of clustering and ordering in multicomponent solid solutions*-I. *Stability criteria*, J. Phys. Chem. Solids, 33 (1972), pp. 297–310.

[14] ———, *An analysis of clustering and ordering in multicomponent solid solutions*-II. *Fluctuations and kinetics*, J. Phys. Chem. Solids, 34 (1973), pp. 1285–1304.

[15] T. DLOTKO, *Fourth order semilinear parabolic equations*, Tsukuba J. Math., 16 (1992), pp. 389–406.

[16] C. M. ELLIOTT, *The Cahn-Hilliard model for the kinetics of phase separation*, in Mathematical Models for Phase Change Problems, J. F. Rodrigues, ed., Birkhäuser-Verlag, Basel, 1989, pp. 35–73.

[17] C. M. ELLIOTT AND D. A. FRENCH, *Numerical studies of the Cahn-Hilliard equation for phase separation*, IMA J. Appl. Math., 38 (1987), pp. 97–128.

[18] C. M. ELLIOTT AND S. LUCKHAUS, *A generalised diffusion equation for phase separation of a multi-component mixture with interfacial free energy*, preprint.

[19] C. M. ELLIOTT AND S. ZHENG, *On the Cahn-Hilliard equation*, Arch. Rational Mech. Anal., 96 (1986), pp. 339–357.

[20] D. J. EYRE, *Systems of Cahn-Hilliard equations*, SIAM J. Appl. Math., 53 (1993), pp. 1686–1712.

[21] G. B. FOLLAND, *Introduction to Partial Differential Equations*, Mathematical Notes, Vol. 17, Princeton University Press, Princeton, NJ, 1976.

[22] C. P. GRANT, *Spinodal decomposition for the Cahn-Hilliard equation*, Comm. Partial Differential Equations, 18 (1993), pp. 453–490.

[23] P. HARTMAN, *Ordinary Differential Equations*, second ed., Birkhäuser-Verlag, Boston, 1982.

[24] J. E. MORRAL AND J. W. CAHN, *Spinodal decomposition in ternary systems*, Acta Metall., 19 (1971), pp. 1037–1045.

[25] K. PROMISLOW, *Time analyticity and Gevrey regularity for solutions of a class of dissipative partial differential equations*, Nonlinear Anal., 16 (1991), pp. 959–980.

[26] S. M. RANKIN, *Semilinear evolution equations in Banach spaces with application to parabolic partial differential equations*, Trans. Amer. Math. Soc., 336 (1993), pp. 523–536.

[27] P. STERNBERG, *The effect of a singular perturbation on nonconvex variational problems*, Arch. Rational Mech. Anal., 101 (1988), pp. 209–260.

[28] R. TEMAM, *Infinite-dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci., Vol. 68, Springer-Verlag, New York, 1988.

[29] J. D. VAN DER WAALS, *The thermodynamic theory of capillarity flow under the hypothesis of a continuous variation in density*, Verh. Konink. Akad. Wetensch. Amsterdam, 1 (1893), pp. 1–56.

[30] S. ZHENG, *Asymptotic behavior of solution to the Cahn-Hilliard equation*, Appl. Anal., 23 (1986), pp. 165–184.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 35–76, January 1995

003

ON A SYSTEM OF INTEGRODIFFERENCE EQUATIONS MODELLING THE PROPAGATION OF GENES*

HWEI-TING LIN†

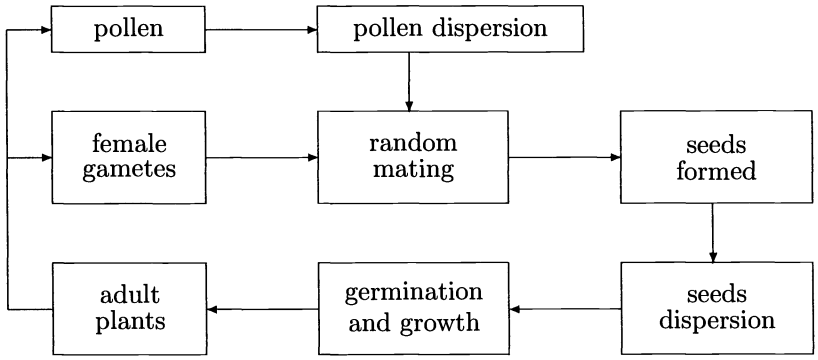
Abstract. The author considers the gene spread of a certain diploid plant, one of whose gene loci has two alleles. A mathematical model is formulated as a system of integrodifference equations $[N^{(n+1)}, u^{(n+1)}] = Q[N^{(n)}, u^{(n)}]$, where Q is a system of nonlinear convolution integral operators. The dynamics of Q in the heterozygote intermediate case is considered in a one-dimensional habitat. When the initial disadvantageous gene is present only in a finite region of the habitat, the dynamics of Q tends to a constant equilibrium dominated by the advantageous gene. Under certain conditions, the system has traveling wave solutions with speed not less than a minimal wave speed. The method of the Laplace transform is used to determine the minimal wave speed explicitly. The Contraction Mapping Theorem is applied to prove the existence and the uniqueness of the traveling waves. Moreover, it is proved that the minimal wave speed is the asymptotic speed of propagation of the advantageous gene, and the dynamics of the model approaches the traveling waves from suitable initial generations.

Key words. traveling waves, asymptotic stability, integrodifference equations, gene spread

AMS subject classifications. 45G10, 92A12, 92A15

1. Introduction.

We study a discrete dynamical system that models the propagation of mutant genes. This model was proposed in a project by Kareiva et al. [6]. They consider a certain diploid plant, one of whose gene loci has two allelic types: A and a. The pollen disperses throughout the habitat according to some probability distribution function $\mathcal{P}$. Then the pollen randomly mates with female gametes. The seeds thus formed disperse throughout the habitat according to some probability distribution function $\mathcal{S}$. The seeds then settle down in the habitat to germinate and grow into adult plants. The new generation of the adult plants starts another cycle of the spread of alleles A and a. See the following flowchart of this model. We copy this flowchart from Kareiva's project [6].



The complete cycle of the gene spread.

We consider the spread of genes in a one-dimensional habitat. Let $u_1(x)$ and $v_1(x)$ denote, respectively, the local densities of female gametes of types a and A in a certain generation. We assume the local densities of pollen of types a and A are given

* Received by the editors September 14, 1992; accepted for publication (in revised form) May 20, 1993.
† Department of Mathematics, Soo Chow University, Taipei 11102, Taiwan.

by ku_1 and kv_1, where k is a constant. The pollen disperses according to a dispersion probability distribution $\mathcal{P}(x)$. Thus the pollen has the densities $ku_1 * \mathcal{P}$ and $kv_1 * \mathcal{P}$ after dispersion, where

$$u_1 * \mathcal{P}(x) = \int_{-\infty}^{\infty} u_1(x-y)d\mathcal{P}(y), \qquad v_1 * \mathcal{P}(x) = \int_{-\infty}^{\infty} v_1(x-y)d\mathcal{P}(y).$$

Then the pollen randomly mates with the female gametes, and a fraction k' of the female gametes produces seeds.

The diploid seeds of genotypes aa, Aa, and AA thus formed have local densities

$$k'u_1(x)\frac{u_1 * \mathcal{P}(x)}{(u_1+v_1) * \mathcal{P}(x)}, \qquad k'\frac{u_1(x)\, v_1 * \mathcal{P}(x) + v_1(x)\, u_1 * \mathcal{P}(x)}{(u_1+v_1) * \mathcal{P}(x)},$$
$$k'v_1(x)\frac{v_1 * \mathcal{P}(x)}{(u_1+v_1) * \mathcal{P}(x)},$$

respectively. These seeds disperse throughout the habitat according to the dispersion probability distribution $\mathcal{S}(x)$. They then settle down and a fraction d of each genotype grows into adult plants. Thus the adult plants of genotype aa, Aa, and AA have local densities

$$k'd\left(u_1\frac{u_1 * \mathcal{P}}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}(x), \qquad k'd\left(\frac{u_1(v_1 * \mathcal{P}) + v_1(u_1 * \mathcal{P})}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}(x),$$
$$k'd\left(v_1\frac{v_1 * \mathcal{P}}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}(x),$$

respectively. We assume that dk' is a constant that does not depend on the genotypes. Therefore the local population density of adults in the generation is given by

$$N_1(x) = dk'\left(u_1\frac{u_1 * \mathcal{P}}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}(x) + dk'\left(\frac{u_1(v_1 * \mathcal{P}) + v_1(u_1 * \mathcal{P})}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}(x)$$
$$+dk'\left(v_1\frac{v_1 * \mathcal{P}}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}(x) \;=\; dk'\,(u_1+v_1) * \mathcal{S}(x).$$

We now assume that each adult plant of type aa produces $\alpha_1(N_1)$ female gametes of type a, each adult plant of type Aa produces $\frac{1}{2}\beta_1(N_1)$ female gametes of type A and the same number of type a, and each adult plant of type AA produces $\gamma_1(N_1)$ female gametes of type A. α_1, β_1, and γ_1 are assumed to depend only on the population density $N_1 = dk'(u_1+v_1)*\mathcal{S}$. Then the densities $\tilde{u}_1(x)$ and $\tilde{v}_1(x)$ of the new generation of female gametes are given by

$$\begin{aligned}
\tilde{u_1} &= \alpha_1(dk'(u_1+v_1) * \mathcal{S})\, dk'\left[\left(u_1\frac{u_1 * \mathcal{P}}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}\right] \\
&\quad +\frac{1}{2}\beta_1(dk'(u_1+v_1) * \mathcal{S})\, dk'\left[\left(\frac{u_1(v_1 * \mathcal{P}) + v_1(u_1 * \mathcal{P})}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}\right], \\
\tilde{v_1} &= \gamma_1(dk'(u_1+v_1) * \mathcal{S})\, dk'\left[\left(v_1\frac{v_1 * \mathcal{P}}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}\right] \\
&\quad +\frac{1}{2}\beta_1(dk'(u_1+v_1) * \mathcal{S})\, dk'\left[\left(\frac{u_1(v_1 * \mathcal{P}) + v_1(u_1 * \mathcal{P})}{(u_1+v_1) * \mathcal{P}}\right) * \mathcal{S}\right].
\end{aligned} \tag{1}$$

The problem is to determine how the genotypes spread in the habitat in the course of many generations. In this paper, we consider a few aspects of this problem in a one-dimensional habitat.

To simplify the equation, we define

$$u_2 = dk'u_1, \quad v_2 = dk'v_1, \quad \alpha_2 = dk'\alpha_1, \quad \beta_2 = dk'\beta_1, \quad \gamma_2 = dk'\gamma_1.$$

Then u_2, v_2 satisfy the same system (1) with α_2, β_2, γ_2 as viability functions and with the constant dk' replaced by 1.

We now assume that there is a positive constant N_0 such that

$$0 \le N\alpha_2(N) \le N_0, \quad 0 \le N\beta_2(N) \le N_0, \quad 0 \le N\gamma_2(N) \le N_0 \tag{2}$$

for $0 \le N \le N_0$. Then we *claim* that, if $0 \le u_2 + v_2 \le N_0$, we also have $0 \le \tilde{u_2} + \tilde{v_2} \le N_0$. To see the claim, we note that

$$\begin{aligned}(\tilde{u_2} + \tilde{v_2})\,[(u_2 + v_2) * \mathcal{S}] \\ \le N_0 \left\{ \left(u_2 \frac{u_2 * \mathcal{P}}{(u_2 + v_2) * \mathcal{P}} \right) * \mathcal{S} + \left(\frac{u_2(v_2 * \mathcal{P}) + v_2(u_2 * \mathcal{P})}{(u_2 + v_2) * \mathcal{P}} \right) * \mathcal{S} \right. \\ \left. + \left(v_2 \frac{v_2 * \mathcal{P}}{(u_2 + v_2) * \mathcal{P}} \right) * \mathcal{S} \right\} = N_0[(u_2 + v_2) * \mathcal{S}].\end{aligned}$$

If $(u_2 + v_2) * \mathcal{S}(x) \neq 0$, then we have $\tilde{u}_2(x) + \tilde{v}_2(x) \le N_0$. If $(u_2 + v_2) * \mathcal{S}(x) = 0$ then

$$\tilde{u}_2(x) + \tilde{v}_2(x) \le \max\{\alpha_2(0), \beta_2(0), \gamma_2(0)\}(u_2 + v_2) * \mathcal{S}(x) = 0.$$

This proves the claim.

We shall assume from now on that the initial gamete population density $u_2 + v_2$, and hence all later densities, satisfy $u_2 + v_2 \le N_0$.

For the convenience of later presentation, we introduce N_0 as a unit of biomass, and renormalize the variables by the following definitions:

$$u = \frac{u_2}{N_0}, \; v = \frac{v_2}{N_0}, \quad \alpha(N) = \alpha_2(N_0 N), \; \beta(N) = \beta_2(N_0 N), \; \gamma(N) = \gamma_2(N_0 N)$$

for $0 \le N \le 1$. Then we still have the system (1) with $k'd = 1$, and the inequalities (2) with $N_0 = 1$. Throughout this paper, we shall use $N(x) = u(x) + v(x) = dk' \frac{N_1(x)}{N_0}$ as the variable of population size. Then $[N, u]$ will be used as our state variables instead of $[u, v]$.

We can now write down the complete cycle as the mapping Q defined in the phase space

$$\Delta = \{[N, u] \mid N, u \in \mathcal{C}(\mathbb{R}), \; 0 \le u(x) \le N(x) \le 1 \text{ for all } x \in \mathbb{R}\},$$

where

$$\mathcal{C}(\mathbb{R}) = \{f \mid f \text{ is a bounded continuous function in } \mathbb{R}.\ f(\pm\infty) \text{ exist}\}.$$

The mapping Q is given by

$$Q[N, u] = [Q_1[N, u], Q_2[N, u]] = [\tilde{N}, \tilde{u}], \qquad [N, u] \in \Delta,$$

where

$$\begin{aligned}\tilde{N} &= Q_1[N,u] \\ &= \alpha(N*\mathcal{S})\left[\left(u\frac{u*\mathcal{P}}{N*\mathcal{P}}\right)*\mathcal{S}\right]+\beta(N*\mathcal{S})\left[\left(u\frac{v*\mathcal{P}}{N*\mathcal{P}}+v\frac{u*\mathcal{P}}{N*\mathcal{P}}\right)*\mathcal{S}\right] \\ &\quad +\gamma(N*\mathcal{S})\left[\left(v\frac{v*\mathcal{P}}{N*\mathcal{P}}\right)*\mathcal{S}\right], \\ \tilde{u} &= Q_2[N,u] \\ &= \alpha(N*\mathcal{S})\left[\left(u\frac{u*\mathcal{P}}{N*\mathcal{P}}\right)*\mathcal{S}\right]+\frac{1}{2}\beta(N*\mathcal{S})\left[\left(u\frac{v*\mathcal{P}}{N*\mathcal{P}}+v\frac{u*\mathcal{P}}{N*\mathcal{P}}\right)*\mathcal{S}\right].\end{aligned}$$

$v = N - u$ in the above equations.

We assume that zero is a point of increase of $\mathcal{P}$ (i.e., $\mathcal{P}(\epsilon) - \mathcal{P}(-\epsilon) > 0$ for all $\epsilon > 0$; see Feller [4]). Thus $N * \mathcal{P}(x) = 0$ will imply $N(x) = 0$. Then we adopt the following convention:

$$u(x)\frac{u*\mathcal{P}(x)}{N*\mathcal{P}(x)} = 0, \quad u(x)\frac{v*\mathcal{P}(x)}{N*\mathcal{P}(x)} = 0, \quad v(x)\frac{u*\mathcal{P}(x)}{N*\mathcal{P}(x)} = 0, \quad v(x)\frac{v*\mathcal{P}(x)}{N*\mathcal{P}(x)} = 0,$$

whenever $N * \mathcal{P}(x) = 0$. Hence Q is well defined in Δ.

The space $\mathcal{C}(\mathbb{R})$ is endowed with the usual sup norm

$$\|f\|_\infty = \sup_{x\in\mathbb{R}} |f(x)|, \qquad f \in \mathcal{C}(\mathbb{R}).$$

We also use the sup norm in Δ unless otherwise stated. Since $\mathcal{P}$ and $\mathcal{S}$ are probability distributions, $f * \mathcal{P} \in \mathcal{C}(\mathbb{R})$ and $f * \mathcal{S} \in \mathcal{C}(\mathbb{R})$ whenever $f \in \mathcal{C}(\mathbb{R})$. This follows from the Lebesgue Dominated Convergence Theorem. Therefore Q maps Δ into itself.

An element of Δ will be called a state of Q. A constant state of Q is a state which is independent of $x \in \mathbb{R}$. Q obviously maps a constant state into a constant state. The corresponding mapping of Q on constant states is the classical one-locus two-allele selection model (see Karlin [7]). α, β, and γ play the roles of the viability fitness matrix elements. However, in contrast with the classical model, α, β, and γ are density dependent. The dynamics of this density-dependent model is presented in Lin [10]. If an advantageous gene is present in a certain generation, the distribution of gene frequencies tends monotonically to the distribution dominated by the advantageous gene, and chaos generically occurs in the evolution of population size due to the density dependence of viability functions.

The modeling of the propagation of an advantageous mutant gene was first introduced by Fisher [5]. He proposed that evolution of the gene frequency $v(x,t)$ of an advantageous gene in a population living in a homogeneous one-dimensional habitat obeys the following nonlinear diffusion equation:

$$v_t = v_{xx} + v(1-v).$$

Fisher [5] and Kolmogorov, Petrovsky, and Piscunoff [8] showed that this model admits traveling waves with speed not less than a minimal speed (which is 2). The wave with minimal speed is an asymptotic state of $v(x,t)$ evolving from some suitable initial conditions. Later work of Aronson and Weinberger [1] went on to show that the minimal wave speed is indeed the asymptotic speed of propagation of the advantageous gene. They also reexamined the modeling process of Fisher's equation, and improved it to take care of both biological realism and mathematical rigor.

One of the recent contributions to the theory of gene spread was a model introduced by Weinberger [15]. He proposed a scalar integrodifference equation of the form

$$v_{n+1}(x) = \int k(x-y)g(v_n(y))dy,$$

where $v_n(x)$ is the gene frequency of the advantageous gene at nth generation, and $k(x)$ is a probability density describing the diffusion process. He showed in [15] that a minimal wave speed c^* exists as

$$c^* = \inf_{\lambda>0}\left\{\frac{1}{\lambda}\log[g'(0)\int e^{\lambda x}k(x)dx]\right\},$$

and it is the asymptotic speed of propagation. He also discussed the connection between his model and Fisher's equation.

In our model, the population density $N(x)$ varies with time and space position. This model is an extension of the model in Weinberger [15], where the population density is assumed to be independent of space position (and hence independent of time). We prove in this paper that a slow selection process enables an advantageous mutant gene to spread throughout the habitat in an asymptotic speed of propagation given by Weinberger's formula. We shall rely on some results in Weinberger [15]. The difficulty in our model is the loss of monotonicity in the system. The use of the comparison principle as presented in Weinberger [16] is limited. The main technique we use is the asymptotic behavior of the solutions of some convolution integral equations as presented in Diekmann and Kaper [2] and Lui [13].

In this paper, $Q^{(n)} = Q \circ \cdots \circ Q$ (n times) denotes the nth iteration of Q. If $[N^{(0)}, u^{(0)}] \in \Delta$ is an initial generation, then $Q^{(n)}[N^{(0)}, u^{(0)}]$ is called the nth iteration state of $[N^{(0)}, u^{(0)}]$ under Q, denoted by $[N^{(n)}, u^{(n)}]$.

2. Basic assumptions and the main results. We state the assumptions for the viability functions and the probability distributions.

(A1) $N\alpha(N)$, $N\beta(N)$, and $N\gamma(N)$ are smooth *concave* functions mapping the interval $0 \le N \le 1$ into itself and $\max\{\alpha(1), \beta(1), \gamma(1)\} < 1$.

(A2) $\alpha(N) \le \beta(N) \le \gamma(N)$ for $0 \le N \le 1$. $\alpha(N)$, $\beta(N)$, and $\gamma(N)$ are not all equal for all $0 \le N < 1$.

(A3) The support of $d\mathcal{S}$ contains a neighborhood of 0, and 0 is a point of increase of $\mathcal{P}$.

The concavity of $N\alpha(N)$ implies that

$$N\frac{d^2}{dN^2}(N\alpha(N)) = N^2\alpha''(N) + 2N\alpha'(N) = \frac{d}{dN}(N^2\alpha'(N)) < 0.$$

Thus $\alpha(N)$, $\beta(N)$, and $\gamma(N)$ are strictly decreasing functions of N. The assumption (A2) is the *heterozygote intermediate* assumption. It shows that the allele A is the advantageous gene. It is then easy to see that

$$Q_1[N,u] \le \gamma(N * \mathcal{S})N * \mathcal{S}$$

for all $[N,u] \in \Delta$. Thus, when $\gamma(0) \le 1$, the dynamics of Q is trivial in the sense that

$$\lim_{n\to\infty} Q^{(n)}[N,u] = [0,0]$$

uniformly in $\mathbb{R}$ for all $[N,u] \in \Delta$. We will always assume the following.

(A4) $\gamma(0) > 1$ and $\gamma'(N^*)N^* + 1 \geq 0$, where N^* satisfies $\gamma(N) = 1$.

Note that N^* is the unique fixed point of the mapping $N\gamma(N)$. It is assumed by (A4) to be a stable fixed point of $N\gamma(N)$ such that any initial condition tends to N^* monotonically under the iteration of $N\gamma(N)$. $[N^*, 0]$ is then a constant fixed point of Q. A straightforward computation shows that the differential of Q at $[N^*, 0]$ is

$$DQ[T, U] = \begin{bmatrix} (\gamma'(N^*)N^* + 1)T * \mathcal{S} + 2(\beta(N^*) - 1)U * \mathcal{J} \\ \beta(N^*)U * \mathcal{J} \end{bmatrix}^t,$$

where $\mathcal{J} = \frac{1}{2}(\mathcal{S} + \mathcal{P} * \mathcal{S})$ and $T,\ U \in \mathcal{C}(\mathbb{R})$. Here the upper index t denotes the transpose operation. Since $\beta(N^*) \leq \gamma(N^*) = 1$ (by (A2)), the Contraction Mapping Theorem shows that the bounded linear operator $DQ : \mathcal{C}(\mathbb{R}) \times \mathcal{C}(\mathbb{R}) \to \mathcal{C}(\mathbb{R}) \times \mathcal{C}(\mathbb{R})$ has a spectrum lying inside the unit disc of the complex plane. Hence $[N^*, 0]$ is a (locally) stable fixed point of the mapping Q whenever (A4) holds. The following result states that $[N^*, 0]$ is in a sense a globally stable fixed point of Q.

THEOREM 2.1. *Let* $[N^{(0)}, u^{(0)}] \in \Delta$ *satisfy* $u^{(0)}(\pm\infty) < N^{(0)}(\pm\infty)$. *Then, under the assumptions* (A1)–(A4), *we have*

$$\lim_{n\to\infty} u^{(n)}(x) = 0, \qquad \lim_{n\to\infty} N^{(n)}(x) = N^*$$

uniformly in $x \in \mathbb{R}$. *Indeed, the rate of convergence is exponential.*

There remains to consider the case where the initial state $[N^{(0)}, u^{(0)}]$ satisfies the equality $u^{(0)}(\pm\infty) = N^{(0)}(\pm\infty)$. We present in this paper several results related to such a case.

We shall impose more assumptions on the viability functions and the distribution functions.

(A2)′ $\alpha(N) < \beta(N) = \gamma(N)$ for $0 \leq N < 1$.

(A3)′ $d\mathcal{P} = \delta$ (the Dirac delta measure). $\mathcal{S}$ has a bounded probability density over $\mathbb{R}$ such that 0 is a point of increase for $\mathcal{S}$ and the Laplace transform

$$K(\lambda) = \int_{-\infty}^{\infty} e^{\lambda x} d\mathcal{S}(x)$$

exists for all $\lambda \in \mathbb{R}$.

(A4)′ $\alpha(0) > 1$ and $\gamma'(N^*)N^* + 1 > 0$.

The assumptions (A2)′ and (A3)′ will simplify the system. The drastic assumption $d\mathcal{P} = \delta$ means that the pollen does not disperse. Although we believe that the following theorems hold even if $d\mathcal{P} \neq \delta$, we are not able to prove them without the restriction. From (A4)′, there exists a unique N_1^* such that $\alpha(N_1^*) = 1$. Then $[N_1^*, N_1^*]$ is also a constant fixed point of Q. It is an unstable constant fixed point of Q. By (A2)′, the strict inequality $N_1^* < N^*$ holds. Define

$$c_1^* = \inf_{\lambda > 0} \frac{1}{\lambda} \log(\gamma(N_1^*)K(\lambda)).$$

The existence of c_1^* is discussed in Weinberger [15], [16]. We then show that c_1^* is the minimal wave speed for a family of traveling waves when the selection is slow.

THEOREM 2.2. *We define* $k(N)$ *such that* $\alpha(N) = k(N)\gamma(N)$ *for* $0 \leq N < 1$. *Then, under assumptions* (A1) *and* (A2)′–(A4)′, *there exist* $\eta > 0$ *such that, whenever* $\max_{0 \leq N \leq N^*}\{|1 - k(N)|, |k'(N)|\} \leq \eta$, *the system* Q *admits a traveling wave profile* $[N(x), u(x)]$ *with speed* c, *which satisfies the system*

$$Q[N, u](x) = [N(x - c), u(x - c)], \qquad x \in \mathbb{R},$$
$$[N(-\infty), u(-\infty)] = [N^*, 0], \qquad [N(\infty), u(\infty)] = [N_1^*, N_1^*]$$

iff $c \geq c_1^*$. *When* $c > c_1^*$, *the wave profile is unique up to translation and is characterized by the positive limit* $\lim_{x\to\infty}(N(x) - N_1^*)e^{\sigma x}$, *where* σ *is the smallest positive root of* $\gamma(N_1^*)e^{-\lambda c}K(\lambda) = 1$.

Furthermore, these waves are the asymptotic state of some initial generation under the evolution of Q. Let us define

$$\Psi(\lambda) = \frac{1}{\lambda}\log(\gamma(N_1^*)K(\lambda)), \qquad \lambda > 0.$$

It is shown in Lemma 4.1 of Weinberger [15, p. 63] that there exists a unique positive value λ^* such that $\Psi(\lambda^*) = c_1^*$. For any $0 < \sigma < \lambda^*$, the value $c = \Psi(\sigma) > c_1^*$ and σ is the smallest positive root of $\gamma(N_1^*)e^{-\lambda c}K(\lambda) = 1$.

THEOREM 2.3. *Under the same assumptions as in Theorem* 2.2, *there exists* $\eta > 0$ *such that the mapping* Q *with the condition* $\max_{0\leq N\leq N^*}\{|1 - k(N)|, |k'(N)|\} \leq \eta$ *has the following asymptotic behavior. Let* $[N^{(0)}(x), u^{(0)}(x)] \in \Delta$ *satisfy* $N^{(0)}(-\infty) > 0$ *and have the asymptotic form for* $x \to \infty$

$$N^{(0)}(x) = L^{(0)} + e^{-\sigma x}\{A^{(0)} + e^{-\epsilon x}R^{(0)}(x)\},$$
$$u^{(0)}(x) = L^{(0)} + e^{-\sigma x}\{C^{(0)} + e^{-\epsilon x}U^{(0)}(x)\}$$

for some $0 < \epsilon < \sigma$, *where* $R^{(0)}(x)$ *and* $U^{(0)}(x)$ *are bounded continuous functions,* $L^{(0)}$, $A^{(0)}$, *and* $C^{(0)}$ *are constants satisfying* $0 < L^{(0)} < 1$, *and* $A^{(0)} \geq C^{(0)}$.

(a) *When* $\sigma \geq \lambda^*$, *then*

$$\lim_{n\to\infty}\max_{x\geq nc}|N^{(n)}(x) - N_1^*| = \lim_{n\to\infty}\max_{x\geq nc}|u^{(n)}(x) - N_1^*| = 0$$

for all $c > c_1^*$.

(b) *When* $0 < \sigma < \lambda^*$ *and* $A^{(0)} = C^{(0)}$, *then*

$$\lim_{n\to\infty}\sup_{x\geq l}|N^{(n)}(x + nc) - N_1^*| = \lim_{n\to\infty}\sup_{x\geq l}|u^{(n)}(x + nc) - N_1^*| = 0,$$

where $c = \Psi(\sigma)$ *and* $l \in \mathbb{R}$ *is arbitrary. Thus*

$$\lim_{n\to\infty}\max_{x\geq nc'}|N^{(n)}(x) - N_1^*| = \lim_{n\to\infty}\max_{x\geq nc'}|u^{(n)}(x) - N_1^*| = 0$$

for all $c' \geq \Psi(\sigma)$.

(c) *When* $0 < \sigma < \lambda^*$ *and* $A^{(0)} > C^{(0)}$, *then*

$$\lim_{n\to\infty}\sup_{x\geq l}|N^{(n)}(x + nc) - N(x)| = \lim_{n\to\infty}\sup_{x\geq l}|u^{(n)}(x + nc) - u(x)| = 0,$$

where $c = \Psi(\sigma)$ *and* $l \in \mathbb{R}$ *is arbitrary.* $[N(x), u(x)]$ *is the traveling wave of* Q *with speed* c, *which is uniquely defined by the limit*

$$\lim_{x\to\infty}(N(x) - N_1^*)e^{\sigma x} = \frac{2(\gamma(N_1^*) - 1)L^{(\infty)}}{\gamma(N_1^*) - (\alpha'(N_1^*)N_1^* + 1)}(A^{(0)} - C^{(0)}),$$

where $L^{(\infty)} = \prod_{n=0}^{\infty}[e^{-\sigma c}K(\sigma)\gamma(L^{(n)})]$ *and* $L^{(n)}$ *is defined by the recursive relation* $L^{(n+1)} = \alpha(L^{(n)})L^{(n)}$ *for* $n = 0, 1, 2, \ldots$. *Thus*

$$\lim_{n\to\infty}\max_{x\geq nc'}|N^{(n)}(x) - N_1^*| = \lim_{n\to\infty}\max_{x\geq nc'}|u^{(n)}(x) - N_1^*| = 0$$

for all $c' > \Psi(\sigma)$.

Theorem 2.3 extends Theorem 1 of Weinberger [15, p. 60] to the system Q. It shows that c_1^* is an asymptotic speed for the spread of the advantageous allele A. Note that the initial state $[N^{(0)}, u^{(0)}] \in \Delta$ in Theorem 2.3 satisfies $N^{(0)}(\infty) = u^{(0)}(\infty) = L^{(0)}$. We will touch upon the cases where $[N^{(0)}(\pm\infty), u^{(0)}(\pm\infty)] = [0, 0]$ in §7.

We now indicate the organization of this paper. Theorem 2.1 is proved in §3. Section 4 covers a key lemma on the asymptotic behavior of the bounded solutions to a linear convolution equation. The techniques used to show such an asymptotic behavior are contained in Diekmann and Kaper [2] and Lui [11]. In §5, we apply the asymptotic representation from §4 along with a (pointwise) Lipschitz estimate of the mapping Q to prove Theorem 2.2 via the Contraction Mapping Theorem. Then Theorem 2.3 is presented in §6 as a direct consequence of the proof of the existence theorem in §5. Finally, in §7, we give some concluding remarks on other traveling wave solutions of Q and their relations to the scalar mapping discussed in Weinberger [15], [16] and Lui [11], [12].

3. The proof of Theorem 2.1. We introduce the gene frequencies

$$m(x) = \frac{u(x)}{N(x)}, \qquad \bar{m}(x) = \frac{u * \mathcal{P}(x)}{N * \mathcal{P}(x)} = \frac{[Nm] * \mathcal{P}(x)}{N * \mathcal{P}(x)}.$$

We rewrite the mapping Q in terms of the variables $[N, m]$. Thus

$$\begin{aligned}\tilde{N} &= \alpha(N * \mathcal{S})\,[Nm\bar{m}] * \mathcal{S} + \beta(N * \mathcal{S})\,[N(m(1-\bar{m}) + \bar{m}(1-m))] * \mathcal{S} \\ &\quad + \gamma(N * \mathcal{S})\,[N(1-m)(1-\bar{m})] * \mathcal{S}, \\ \tilde{m} &= \frac{\alpha(N * \mathcal{S})\,[Nm\bar{m}] * \mathcal{S} + \frac{1}{2}\beta(N * \mathcal{S})\,[N(m(1-\bar{m}) + \bar{m}(1-m))] * \mathcal{S}}{\tilde{N}}.\end{aligned}$$

For each $[N(x), m(x)]$, we define two functions:

$$\begin{aligned}p(x,y) &= \frac{1}{2}N(y)\{\alpha(N * \mathcal{S}(x))[m(y) + \bar{m}(y)] \\ &\qquad + \beta(N * \mathcal{S}(x))[(1 - m(y)) + (1 - \bar{m}(y))]\}, \\ q(x,y) &= N(y)\{\alpha(N * \mathcal{S}(x))m(y)\bar{m}(y) + \beta(N * \mathcal{S}(x))[m(y)(1 - \bar{m}(y)) \\ &\qquad + (1 - m(y))\bar{m}(y)] + \gamma(N * \mathcal{S}(x))(1 - m(y))(1 - \bar{m}(y))\}.\end{aligned}$$

LEMMA 3.1. *Assume that* $\alpha(N) \le \beta(N) \le \gamma(N)$ *for all* $0 \le N < 1$. *Let* $N(x)$, $m(x)$ *be two continuous functions in* $\mathbb{R}$ *such that* $0 \le N(x)$, $m(x) \le 1$ *for all* $x \in \mathbb{R}$. *Then*

(a) $p(x,y) \le q(x,y)$ *for all* $x, y \in \mathbb{R}$,

(b) $\bar{m}(x) \le \|m\|_\infty (\int_{-\infty}^{\infty} p(x, x-y)d\mathcal{S}(y) / \int_{-\infty}^{\infty} q(x, x-y)d\mathcal{S}(y)) \le \|m\|_\infty$ *for all* $x \in \mathbb{R}$.

Proof. (a)

$$\begin{aligned}&q(x,y) - p(x,y) \\ &= \frac{1}{2}N(y)\{2\alpha m(y)\bar{m}(y) + 2\beta[m(y)(1 - \bar{m}(y)) + (1 - m(y))\bar{m}(y)] \\ &\qquad + 2\gamma(1 - m(y))(1 - \bar{m}(y)) - \alpha[m(y) + \bar{m}(y)] \\ &\qquad - \beta[(1 - m(y)) + (1 - \bar{m}(y))]\} \\ &= \frac{1}{2}N(y)\{(1 - \bar{m}(y))[(\beta - \alpha)m(y) + (\gamma - \beta)(1 - m(y))] \\ &\qquad + (1 - m(y))[(\beta - \alpha)\bar{m}(y) + (\gamma - \beta)(1 - \bar{m}(y))]\}.\end{aligned}$$

When $0 \le m(y) \le 1$ for all $y \in \mathbb{R}$, it is clear that $0 \le \bar{m}(y) \le 1$ for all $y \in \mathbb{R}$. Since $\alpha \le \beta \le \gamma$, we get $p(x,y) \le q(x,y)$ for all $x, y \in \mathbb{R}$.

(b) We note that

$$\begin{aligned}
&\alpha(N*\mathcal{S})[Nm\bar{m}]*\mathcal{S} + \frac{1}{2}\beta(N*\mathcal{S})[N(m(1-\bar{m})+\bar{m}(1-m))]*\mathcal{S} \\
&= \frac{1}{2}\{\alpha(N*\mathcal{S})[Nm\bar{m}]*\mathcal{S} + \beta(N*\mathcal{S})[Nm(1-\bar{m})]*\mathcal{S}\} \\
&\quad + \frac{1}{2}\{\alpha(N*\mathcal{S})[Nm\bar{m}]*\mathcal{S} + \beta(N*\mathcal{S})[N\bar{m}(1-m)]*\mathcal{S}\} \\
&\le \frac{1}{2}\|m\|_\infty\{\alpha(N*\mathcal{S})[N\bar{m}]*\mathcal{S} + \beta(N*\mathcal{S})[N(1-\bar{m})]*\mathcal{S}\} \\
&\quad + \frac{1}{2}\|\bar{m}\|_\infty\{\alpha(N*\mathcal{S})[Nm]*\mathcal{S} + \beta(N*\mathcal{S})[N(1-m)]*\mathcal{S}\}.
\end{aligned}$$

By the definition of $\bar{m}(y)$, we have $\|\bar{m}\|_\infty \le \|m\|_\infty$. Therefore

$$\begin{aligned}
&\alpha(N*\mathcal{S})[Nm\bar{m}]*\mathcal{S} + \frac{1}{2}\beta(N*\mathcal{S})[Nm(1-\bar{m}) + N\bar{m}(1-m)]*\mathcal{S} \\
&\le \|m\|_\infty \int_{-\infty}^{\infty} p(x, x-y)d\mathcal{S}(y).
\end{aligned}$$

By the definition of $q(x,y)$, the identity

$$\tilde{N}(x) = \int_{-\infty}^{\infty} q(x, x-y)d\mathcal{S}(y)$$

holds for all $x \in \mathbb{R}$. Thus

$$\tilde{m}(x) \le \|m\|_\infty \frac{\int_{-\infty}^{\infty} p(x, x-y)d\mathcal{S}(y)}{\int_{-\infty}^{\infty} q(x, x-y)d\mathcal{S}(y)}$$

for all $x \in \mathbb{R}$. Then (a) implies (b). □

We start to prove Theorem 2.1. When $m^{(0)}(x) \equiv 0$, $m^{(n)}(x) \equiv 0$ holds for all $n = 1, 2, 3, \ldots$. Then $N^{(n)}(x)$ satisfies

$$N^{(n+1)} = \gamma(N^{(n)} * \mathcal{S})\, N^{(n)} * \mathcal{S}.$$

An application of Theorem 3 of Weinberger [15] proves this theorem. Thus we shall assume that $m^{(0)}(x) \not\equiv 0$.

The rest of the proof will be divided into several steps.

Step 1. $N^{(n)}(\pm\infty) > 0$ and $m^{(n+1)}(\pm\infty) \le m^{(n)}(\pm\infty) < 1$ for each n. Moreover, $\lim_{n\to\infty} m^{(n)}(\pm\infty) = 0$ and $\lim_{n\to\infty} N^{(n)}(\pm\infty) = N^*$.

For convenience, let $r_n = N^{(n)}(\infty)$, $s_n = m^{(n)}(\infty)$. Then $[r_n, s_n]$ satisfies the iteration relations

$$\begin{aligned}
r_{n+1} &= r_n\{\alpha(r_n)s_n^2 + 2\beta(r_n)s_n(1-s_n) + \gamma(r_n)(1-s_n)^2\}, \\
s_{n+1} &= \frac{\alpha(r_n)s_n^2 + \beta(r_n)s_n(1-s_n)}{\alpha(r_n)s_n^2 + 2\beta(r_n)s_n(1-s_n) + \gamma(r_n)(1-s_n)^2}.
\end{aligned}$$

By (A2), we have $r_{n+1} \ge \alpha(r_n)r_n$. Thus $r_n > 0$ by induction. Now

$$s_{n+1} - s_n = s_n(1-s_n)\frac{(\beta(r_n)-\gamma(r_n))(1-s_n) + (\alpha(r_n)-\beta(r_n))s_n}{\alpha(r_n)s_n^2 + 2\beta(r_n)s_n(1-s_n) + \gamma(r_n)(1-s_n)^2}.$$

The assumption (A2) implies that $s_{n+1} \leq s_n \leq s_0 < r_0 \leq 1$. The sequence s_n has a limit. From the iteration relations for $[r_n, s_n]$, it is easy to see that $\lim_{n\to\infty} s_n = 0$ and $\lim_{n\to\infty} r_n = N^*$ (see Lin [10]).

Step 2. $\limsup_{n\to\infty} N^{(n)}(x) \leq N^*$ uniformly in $x \in \mathbb{R}$.

Let $M = \max_{0\leq N\leq 1} N\gamma(N)$. We note that

$$Q_1[N, u] \leq \gamma(N * \mathcal{S})N * \mathcal{S} \leq M.$$

By (A1) and (A4), $N\gamma(N)$ is strictly increasing over $0 \leq N \leq M$. Then an easy induction argument shows that

$$N^{(n)}(x) \leq M_{n-1}, \qquad n = 1, 2, \ldots,$$

where M_n is defined by

$$M_{n+1} = \gamma(M_n)M_n, \quad M_0 = M, \quad n = 0, 1, 2, \ldots.$$

But $\lim_{n\to\infty} M_n = N^*$. This proves Step 2.

Step 3. There exists some integer k such that $N^{(n)}(x) > 0$ and $m^{(n)}(x) < 1$ for all $x \in \mathbb{R}$ and for all $n \geq k$.

We define

$$Z_n = \{x \in \mathbb{R} | N^{(n)}(x) = 0\}, \qquad T_n = \{x \in \mathbb{R} | m^{(n)}(x) = 1\}.$$

Since $N^{(n)}(x)$ and $m^{(n)}(x)$ are continuous, Z_n and T_n are closed subsets of $\mathbb{R}$. By Step 1, Z_n and T_n are bounded subsets of $\mathbb{R}$. Thus they are compact subsets of $\mathbb{R}$. We *claim* that

$$Z_{n+1} \subseteq Z_n, \qquad n = 0, 1, 2, \ldots.$$

Let $x_0 \in Z_{n+1}$. Then $N^{(n+1)}(x_0) = 0$. By

$$N^{(n+1)}(x_0) \geq \alpha(N^{(n)} * \mathcal{S}(x_0))\, N^{(n)} * \mathcal{S}(x_0),$$

we get $N^{(n)} * \mathcal{S}(x_0) = 0$. Thus $N^{(n)}(x_0 - y) = 0$ for all y in the support of $d\mathcal{S}$. Since the support of $d\mathcal{S}$ contains a neighborhood of 0, $N^{(n)}(x_0) = 0$. Thus $x_0 \in Z_n$. This proves the claim. From the claim, we want to deduce that there exists an integer k_1, such that $Z_{k_1} = \emptyset$. Otherwise, $Z_n \neq \emptyset$ for all n. Then $\cap_n Z_n \neq \emptyset$. Let $x_0 \in \cap_n Z_n$. Thus $N^{(n)}(x_0) = 0$ for all n. By the same argument of the proof of the claim, we know that

$$N^{(0)}(x_0 - y_1 - y_2 - \cdots - y_n) = 0$$

for all $y_1, y_2, \ldots, y_n$ in the support of $d\mathcal{S}$. But the support of $d\mathcal{S}$ contains a neighborhood of 0, and we have

$$\cup_n \{x_0 - y_1 - \cdots - y_n | y_1, \ldots, y_n \in \text{ supp}(d\mathcal{S})\} = \mathbb{R}.$$

Therefore $N^{(0)}(x) = 0$ for all x. This contradicts $N^{(0)}(\pm\infty) > 0$. Thus we have proved that $N^{(n)}(x) > 0$ for all $x \in \mathbb{R}$ and $n \geq k_1$.

Now, we *claim* again that

$$T_{n+1} \subseteq T_n, \qquad n = k_1, k_1 + 1, k_1 + 2, \ldots.$$

Let $x_0 \in T_{n+1}$. Then $m^{(n+1)}(x_0) = 1$. From Lemma 3.1 (b), we get

$$1 = m^{(n+1)}(x_0) \le \frac{\int_{-\infty}^{\infty} p^{(n)}(x_0, x_0 - y) d\mathcal{S}(y)}{\int_{-\infty}^{\infty} q^{(n)}(x_0, x_0 - y) d\mathcal{S}(y)},$$

where $p^{(n)}(x, y)$ and $q^{(n)}(x, y)$ are the functions p and q defined by $[N^{(n)}(x), m^{(n)}(x)]$. Since $p^{(n)}(x, y) \le q^{(n)}(x, y)$ for all (x, y), we get

$$\int_{-\infty}^{\infty} [q^{(n)}(x_0, x_0 - y) - p^{(n)}(x_0, x_0 - y)] d\mathcal{S}(y) = 0.$$

Therefore $q^{(n)}(x_0, x_0 - y) - p^{(n)}(x_0, x_0 - y) = 0$ for all y in the support of $d\mathcal{S}$. From the expression for $q^{(n)}(x, y) - p^{(n)}(x, y)$ in the proof of Lemma 3.1 (a) and $N^{(n)}(y) > 0$, we have

$$(1 - m^{(n)}(x_0 - y))\{(\beta - \alpha)\bar{m}^{(n)}(x_0 - y) + (\gamma - \beta)(1 - \bar{m}^{(n)}(x_0 - y))\} = 0.$$

By the assumption (A2), $(\beta - \alpha)\bar{m}^{(n)}(x_0 - y) + (\gamma - \beta)(1 - \bar{m}^{(n)}(x_0 - y)) \neq 0$ for all $y \in \mathbb{R}$. Thus $m^{(n)}(x_0 - y) = 1$ for all y in the support of $d\mathcal{S}$. In particular, we get $m^{(n)}(x_0) = 1$ by putting $y = 0$. This proves the claim. From the claim, we can find $k \ge k_1$ such that $T_k = \emptyset$. Otherwise, if $T_n \neq \emptyset$ for all $n \ge k_1$, then there exists $x_0 \in \cap_{n \ge k_1} T_n$. From $m^{(n)}(x_0) = 1$ for all $n \ge k_1$, we deduce that $m^{(k_1)}(x) = 1$ for all $x \in \mathbb{R}$. This contradicts $m^{(k_1)}(\pm\infty) < 1$. This proves Step 3.

Step 4. $\|m^{(n)}\|_\infty$ decreases to 0.

For convenience, we let $a_n = \|m^{(n)}\|_\infty$. By Lemma 3.1 (b), we have $a_{n+1} \le a_n$. Thus $\lim_{n\to\infty} a_n = a$ exists. By Steps 1 and 3, $0 \le a_n < 1$ for all $n \ge k$. Thus $0 \le a < 1$.

Let us consider the function

$$g(N, m, \bar{m}) = \frac{1}{2} \frac{\alpha(N)(m + \bar{m}) + \beta(N)(2 - m - \bar{m})}{\alpha(N)m\bar{m} + \beta(N)(m + \bar{m} - 2m\bar{m}) + \gamma(N)(1 - m)(1 - \bar{m})},$$

which is defined in $0 \le N,\ m,\ \bar{m} \le 1$. The function g is the ratio of the functions p and q. Note that $g(N, m, \bar{m}) = g(N, \bar{m}, m)$. By the assumption (A2), a computation shows that

$$g(N, m, \bar{m}) \le \max\{g(N, a_n, a_n),\ g(N, a_n, 0),\ g(N, 0, 0)\}$$

for all $0 \le N < 1$, $0 \le m,\ \bar{m} \le a_n$. We define

$$t_n = \max_{0 \le N \le \frac{1}{2}(1+N^*)} \{g(N, a_n, a_n),\ g(N, a_n, 0),\ g(N, 0, 0)\}.$$

Note that, for $0 \le N \le \frac{1}{2}(1 + N^*)$, $g(N, m, \bar{m}) = 1$ only if $m = 1$ or $\bar{m} = 1$. Thus when $0 < a_n < 1$, we have $g(N, m, \bar{m}) \le t_n < 1$ for all $0 \le N \le \frac{1}{2}(1 + N^*)$, $0 \le m,\ \bar{m} \le a_n$. By (A4), Step 2 implies that we can choose k in Step 3 such that $0 < N^{(n)}(x) \le \frac{1}{2}(1 + N^*)$ for all $x \in \mathbb{R}$ and $n \ge k$. Thus, for $n \ge k$, we get

$$g(N^{(n)} * \mathcal{S}(x), m^{(n)}(y), \bar{m}^{(n)}(y)) \le t_n$$

for all $x, y \in \mathbb{R}$. But

$$g(N^{(n)} * \mathcal{S}(x), m^{(n)}(y), \bar{m}^{(n)}(y)) = \frac{p^{(n)}(x, y)}{q^{(n)}(x, y)}.$$

Hence

$$p^{(n)}(x, x-y) \le t_n q^{(n)}(x, x-y)$$

for all $x, y \in \mathbb{R}$. Integrating the above inequality with respect to $d\mathcal{S}(y)$, we get

$$m^{(n+1)}(x) \le \|m^{(n)}\|_\infty \frac{\int_{-\infty}^{\infty} p^{(n)}(x, x-y)d\mathcal{S}(y)}{\int_{-\infty}^{\infty} q^{(n)}(x, x-y)d\mathcal{S}(y)} \le a_n t_n$$

for all $x \in \mathbb{R}$. Therefore $a_{n+1} \le a_n t_n$ for $n \ge k$. We let $n \to \infty$. Then $a \le a\,t$ where

$$t = \max_{0 \le N \le \frac{1}{2}(1+N^*)} \max\{g(N, a, a),\ g(N, a, 0),\ g(N, 0, 0)\}.$$

Since $0 \le a < 1$, we have $0 < t < 1$. Therefore $a = 0$.

We note that this argument shows that $\|m^{(n)}\|_\infty$ decreases to 0 exponentially.

Step 5. Let $b_n = \inf_{x \in \mathbb{R}} N^{(n)}(x)$. Then there exists an $l > 0$ such that $b_n \ge l$ for all $n \ge k$, where k is the integer in Step 3.

We note that, from Steps 1 and 3, $b_n > 0$ for all $n \ge k$. We can rewrite $N^{(n+1)}(x)$ as

$$\begin{aligned} N^{(n+1)}(x) &= \alpha(N^{(n)} * \mathcal{S})\,[N^{(n)} m^{(n)} \bar{m}^{(n)}] * \mathcal{S} \\ &\quad + \beta(N^{(n)} * \mathcal{S})\,[N^{(n)}(m^{(n)}(1-\bar{m}^{(n)}) + \bar{m}^{(n)}(1-m^{(n)}))] * \mathcal{S} \\ &\quad + \gamma(N^{(n)} * \mathcal{S})\,[N^{(n)}(1-m^{(n)})(1-\bar{m}^{(n)})] * \mathcal{S} \\ &\equiv \gamma(N^{(n)} * \mathcal{S}(x))\, N^{(n)} * \mathcal{S}(x) + D_n(x). \end{aligned}$$

Since $\|\bar{m}\|_\infty \le \|m\|_\infty$ and $N^{(n)}(x)$ are uniformly bounded, Step 4 implies that $d_n = \|D_n\|_\infty$ tends to 0 as $n \to \infty$. By (A4), we can choose $\epsilon_0 > 0$ such that $\gamma(N)N \ge \gamma(b)b$ for all $b \le N^*$ and $b \le N \le N^* + \epsilon_0$. Step 2 implies that we can find $k_2 \ge k$ such that $N^{(n)}(x) \le N^* + \epsilon_0$ for all $x \in \mathbb{R}$ and $n \ge k_2$. Therefore, when $n \ge k_2$, $N^{(n+1)}(x) \ge \gamma(b_n)b_n - d_n$ for all $x \in \mathbb{R}$. We deduce that

$$b_{n+1} \ge \gamma(b_n)b_n - d_n \qquad \text{for } n \ge k_2.$$

We *claim* that if there exists a subsequence b_{n_i} such that b_{n_i} has a positive lower bound l, then b_n has a positive lower bound for $n \ge k$. To prove the claim, we may choose the lower bound l such that $0 < l < N^*$. Then $\gamma(l)l - l > 0$. Thus we can choose $k_3 \ge k_2$ such that $d_n < \gamma(l)l - l$ for $n \ge k_3$. We can find an index n_i with $n_i \ge k_3$ and $b_{n_i} \ge l$. Then we will prove by induction that $b_n \ge l$ for $n \ge n_i$. Assume that $b_n \ge l$ for some $n \ge n_i$. Then

$$b_{n+1} \ge \gamma(b_n)b_n - d_n \ge \gamma(l)l - d_n = (\gamma(l)l - l) - d_n + l > l.$$

This proves the claim.

We know that b_n is a bounded sequence. If Step 5 is false, it follows from the claim that $\lim_{n\to\infty} b_n = 0$. From Step 1, $N^{(n)}(\pm\infty)$ tends to N^* as $n \to \infty$. Thus, for each $n \ge k_2$, we can find $x_n \in \mathbb{R}$ such that $N^{(n)}(x_n) = b_n$. This can be done by taking k_2 larger whenever it is necessary. We call

$$e_n = N^{(n)} * \mathcal{S}(x_{n+1}) \quad \text{for } n \ge k_2.$$

Then

$$\begin{aligned} b_{n+1} &= N^{(n+1)}(x_{n+1}) \\ &= \gamma(N^{(n)} * \mathcal{S}(x_{n+1}))\, N^{(n)} * \mathcal{S}(x_{n+1}) + D_n(x_{n+1}) \\ &= \gamma(e_n)e_n + D_n(x_{n+1}), \qquad n \geq k_2. \end{aligned}$$

Since $D_n(x_{n+1}) = O(\|m^{(n)}\|_\infty)$ as $n \to \infty$, we get $\lim_{n\to\infty} \gamma(e_n)e_n = 0$. But $0 \leq N^{(n)} \leq \frac{1}{2}(1 + N^*)$, $\gamma(e_n) \geq \gamma(\frac{1}{2}(1 + N^*)) > 0$. Thus $\lim_{n\to\infty} e_n = 0$. Now, from the iteration relation for $N^{(n+1)}$, we get

$$\begin{aligned} N^{(n+1)}(x) \geq b_n\{&\alpha(N^{(n)} * \mathcal{S})\,[m^{(n)}\bar{m}^{(n)}] * \mathcal{S} \\ &+ \beta(N^{(n)} * \mathcal{S})\,[m^{(n)}(1 - \bar{m}^{(n)}) + \bar{m}^{(n)}(1 - m^{(n)})] * \mathcal{S} \\ &+ \gamma(N^{(n)} * \mathcal{S})\,[(1 - m^{(n)})(1 - \bar{m}^{(n)})] * \mathcal{S}\} \\ \equiv b_n\{&\gamma(N^{(n)} * \mathcal{S}(x)) + \tilde{D}_n(x)\}, \end{aligned}$$

where $\tilde{D}_n(x) = O(\|m^{(n)}\|_\infty)$. $\tilde{D}_n(x)$ tends to 0 uniformly in $x \in \mathbb{R}$. By putting $x = x_{n+1}$ in the inequality, we have

$$b_{n+1} \geq b_n\{\gamma(e_n) + \tilde{D}_n(x_{n+1})\}, \qquad n \geq k_2.$$

However,

$$\lim_{n\to\infty} \{\gamma(e_n) + \tilde{D}_n(x_{n+1})\} = \gamma(0) > 1.$$

We can find $k_4 \geq k_2$ such that $\gamma(e_n) + \tilde{D}_n(x_{n+1}) > 1$ for all $n \geq k_4$. Thus $b_{n+1} > b_n > 0$ for $n \geq k_4$. This is a contradiction to $\lim_{n\to\infty} b_n = 0$. The proof of Step 5 is complete.

Step 6. $\lim_{n\to\infty} b_n = N^*$. Hence $N^{(n)}(x)$ converges to N^* uniformly in $\mathbb{R}$.

Let $\epsilon > 0$ be given sufficiently small. From Steps 2 and 5, we can find $k_2 \geq k$ such that

$$0 < l \leq b_n \leq N^{(n)}(x) \leq N^* + \epsilon, \qquad b_{n+1} \geq \gamma(b_n)b_n - d_n$$

for $n \geq k_2$ and for all $x \in \mathbb{R}$. Define

$$l_1 = \min_{l \leq N \leq N^* - \epsilon} (\gamma(N)N - N),$$

$l_1 > 0$. Thus we can find $k_5 \geq k_2$ such that $d_n < l_1$ for $n \geq k_5$. We *claim* that there exists $k_6 \geq k_5$ such that $b_{k_6} \geq N^* - \epsilon$. We prove the claim by contradiction. Suppose that $b_n < N^* - \epsilon$ for all $n \geq k_5$. From

$$b_{n+1} \geq \gamma(b_n)b_n - d_n = (\gamma(b_n)b_n - b_n) - d_n + b_n \geq l_1 - d_n + b_n > b_n$$

for all $n \geq k_5$, we know that $\lim_{n\to\infty} b_n = b$ exists. Clearly, $l \leq b \leq N^* - \epsilon$. However, by letting $n \to \infty$ in $b_{n+1} \geq b_n\gamma(b_n) - d_n$, we get $b \geq b\gamma(b)$. Thus $\gamma(b) \leq 1$. But $b < N^*$, and we also have $\gamma(b) > 1$. This is a contradiction, and the claim is proved. From the claim, we can prove by induction that $b_n \geq N^* - \epsilon$ for $n \geq k_6$. Suppose that this inequality holds for some n. Then

$$\begin{aligned} b_{n+1} \geq \gamma(b_n)b_n - d_n &\geq \gamma(N^* - \epsilon)(N^* - \epsilon) - d_n \\ &= [\gamma(N^* - \epsilon)(N^* - \epsilon) - (N^* - \epsilon)] - d_n + (N^* - \epsilon) \\ &\geq l_1 - d_n + (N^* - \epsilon) \geq N^* - \epsilon. \end{aligned}$$

Thus we have proved that $|b_n - N^*| \leq \epsilon$ for all $n \geq k_6$.

The proof of Theorem 2.1 is complete. □

Remark 3.1. From the proof of Step 3, we see that if $N^{(0)}(x) \not\equiv 0$ (without assuming $N^{(0)}(\pm\infty) > 0$), then for any compact subset $K \subseteq \mathbb{R}$, there exists n_0 such that $N^{(n)}(x) > 0$ for all $x \in K$ and $n \geq n_0$.

Remark 3.2. If we assume instead of (A3) that each point of $\mathbb{R}$ is a point of increase for $d\mathcal{S}$, then we can obtain the following conclusion. Let $[N^{(0)}, u^{(0)}] \in \Delta$ satisfy

$$N^{(0)}(x) \not\equiv 0 \quad \text{and} \quad m^{(0)}(\pm\infty) = \lim_{x\to\pm\infty} \frac{u^{(0)(x)}}{N^{(0)}(x)} < 1;$$

then, under assumptions (A1), (A2), and (A4), we have $u^{(n)}(x) \to 0$ uniformly in $\mathbb{R}$ and $N^{(n)}(x) \to N^*$ uniformly on compact subsets of $\mathbb{R}$.

4. A preliminary lemma. In this section, we shall present a lemma on the asymptotic behavior of a positive bounded solution $\varphi(x)$ to the linear convolution equation

$$\varphi(x) = f(x)\,\varphi * F(x),$$

where F is a probability distribution function on $\mathbb{R}$ and $f(x)$ is a given continuous function. This lemma is essential to the proof of the existence of the traveling wave solutions of Q. We shall use several results in Essén [3] and Diekmann and Kaper [2]. A full discussion on the existence and the uniqueness of the bounded solutions to the above convolution equation can be found in Lin [9].

We assume that F has a bounded probability density such that the Laplace transform

$$\hat{F}(\lambda) = \int_{-\infty}^{\infty} e^{\lambda x} dF(x)$$

exists for all $\lambda \geq 0$. We define $F_0 = \min_{\lambda\geq 0} \hat{F}(\lambda)$.

LEMMA 4.1. *Let $f \in C(\mathbb{R})$ satisfy $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $f(\infty) > 1$. Suppose that $\varphi \in C(\mathbb{R})$ satisfies $\varphi(x) \geq 0$, $\varphi(x) \not\equiv 0$, and*

$$\varphi(x) = f(x)\,\varphi * F(x) \quad \textit{for all } x \in \mathbb{R}.$$

We make the following assertions.

(a) *There exists some $k > 0$ such that $\int_{-\infty}^{\infty} e^{\lambda x}\varphi(x)dx < \infty$ for $0 < \lambda < k$.*

(b) *Define $\sigma = \sup\{\lambda | \lambda > 0, \int_{-\infty}^{\infty} e^{\lambda x}\varphi(x)dx < \infty\}$. Then $\varphi(x)e^{\lambda x}$ is uniformly bounded for all $0 \leq \lambda < \sigma$, while $\varphi(x)e^{\lambda x}$ is unbounded for $\lambda > \sigma$.*

(c) *If we assume further that $\frac{1}{F_0} > f(\infty) \geq f(x)$, $f(x)$ is not a constant function, and $f(x) = f(\infty) + O(e^{-\epsilon x})$ as $x \to \infty$ for some $\epsilon > 0$, then $\varphi(x)$ has the asymptotic representation*

$$\varphi(x) = e^{-\sigma x}[A + U(x)e^{-\epsilon' x}],$$

where A, ϵ', and σ are some positive constants. $U(x)$ is a bounded continuous function with $Ue^{-\epsilon' x}$ bounded, and $U(x) \leq 0$ for all $x \in \mathbb{R}$. Moreover, σ is the smallest positive root of the equation $f(\infty)\hat{F}(\lambda) = 1$.

Proof. (a). Let $1 < t < f(\infty)$ and choose $l > 0$ such that $F(l) - F(-l) = \frac{1}{2}(1 + \frac{1}{t})$. Define

$$F_l(x) = \begin{cases} 0 & \text{if } x \le -l, \\ \frac{F(x)-F(-l)}{F(l)-F(-l)} & \text{if } -l \le x \le l, \\ 1 & \text{if } l < x. \end{cases}$$

$F_l(x)$ is again a probability distribution. Now

$$\begin{aligned} &\varphi - \varphi * F_l \\ &\quad = \varphi(x) - \frac{2t}{1+t}\int_{-l}^{l} \varphi(x-y)dF(y) \\ &\quad = \frac{t-1}{t+1}\varphi(x) + \frac{2}{t+1}\varphi(x) - \frac{2t}{t+1}\int_{-l}^{l} \varphi(x-y)dF(y) \\ &\quad = \frac{t-1}{t+1}\varphi(x) + \frac{2}{t+1}\left\{ f(x)\int_{-\infty}^{\infty} \varphi(x-y)dF(y) - t\int_{-l}^{l} \varphi(x-y)dF(y) \right\} \\ &\quad = \frac{t-1}{t+1}\varphi(x) + \frac{2}{t+1}\psi(x), \end{aligned}$$

where $\psi(x)$ can be written as

$$\psi(x) = (f(x) - t)\int_{-l}^{l} \varphi(x-y)dF(y) + f(x)\int_{|y|\ge l} \varphi(x-y)dF(y).$$

Choose x_0 such that $f(x) > t$ for $x \ge x_0$. Then $\psi(x) \ge 0$ for $x \ge x_0$. By integrating the above identity from x to r with $x_0 \le x \le r$, we get

$$\varphi * N_l(r) - \varphi * N_l(x) = \frac{t-1}{t+1}\int_x^r \varphi(y)dy + \frac{2}{t+1}\int_x^r \psi(y)dy, \tag{3}$$

where

$$N_l(x) = \begin{cases} 1 - F_l(x) & \text{if } x \ge 0, \\ -F_l(x) & \text{if } x < 0. \end{cases}$$

Note that $\int_{-\infty}^{\infty} |x|dF_l(x) < \infty$ and $\int_{-\infty}^{\infty} |N_l(x)|dx = \int_{-\infty}^{\infty} |x|dF_l(x)$. Since $f(\infty) > 1$ and $\varphi \in \mathcal{C}(\mathbb{R})$, $\varphi(\infty) = 0$. Thus $\lim_{r\to\infty} \varphi * N_l(r) = 0$. Therefore, using $\varphi \ge 0$ and $\psi \ge 0$ in (3), we have proved that $\int_x^{\infty} \varphi(y)dy$ and $\int_x^{\infty} \psi(y)dy$ exist for all $x \ge x_0$, and they have the relation

$$\int_x^{\infty} \varphi(y)dy = -\frac{t+1}{t-1}\left\{ \varphi * N_l(x) + \frac{2}{t+1}\int_x^{\infty} \psi(y)dy \right\}.$$

We shall generalize this relation as follows.

Define $\varphi^{(n)}$ and $\psi^{(n)}$ inductively as

$$\varphi^{(n+1)}(x) = \int_x^{\infty} \varphi^{(n)}(y)dy, \qquad \varphi^{(0)}(x) = \varphi(x),$$

$$\psi^{(n+1)}(x) = \int_x^{\infty} \psi^{(n)}(y)dy, \qquad \psi^{(0)}(x) = \psi(x),$$

for $n = 0, 1, 2, \ldots$. We *claim* that $\varphi^{(n)}$ and $\psi^{(n)}$ exist for all n and they satisfy the following identities

$$\varphi^{(n)} - \varphi^{(n)} * F_l = \frac{t-1}{t+1}\varphi^{(n)} + \frac{2}{t+1}\psi^{(n)}, \tag{4}$$

$$\varphi^{(n+1)}(x) = -\frac{t-1}{t+1}\left[\varphi^{(n)} * N_l(x) + \frac{2}{t+1}\psi^{(n+1)}(x)\right] \tag{5}$$

for $n = 0, 1, 2, \ldots$, and for all $x \geq x_0$. We have proved the case $n = 0$. Assume that (4) and (5) have been proved in step n; then $\varphi^{(n+1)}$ and $\psi^{(n+1)}$ exist. We integrate (4) from x to ∞, and note that

$$\begin{aligned}
\int_x^\infty \varphi^{(n)} * F_l(y)dy &= \int_x^\infty \left(\int_{-\infty}^\infty \varphi^{(n)}(y-z)dF_l(z)\right)dy \\
&= \int_{-\infty}^\infty \left(\int_x^\infty \varphi^{(n)}(y-z)dy\right)dF_l(z) \\
&= \int_{-\infty}^\infty \left(\int_{x-z}^\infty \varphi^{(n)}(y)dy\right)dF_l(z) \\
&= \int_{-\infty}^\infty \varphi^{(n+1)}(x-z)dF_l(z) \\
&= \varphi^{(n+1)} * F_l(x).
\end{aligned}$$

This proves that (4) holds for step $n+1$. We can now integrate (4) for step $n+1$ from x to r. We get

$$\varphi^{(n+1)} * N_l(r) - \varphi^{(n+1)} * N_l(x) = \frac{t-1}{t+1}\int_x^r \varphi^{(n+1)}(y)dy + \frac{2}{t+1}\int_x^r \psi^{(n+1)}(y)dy.$$

By definition, $\varphi^{(n+1)}(\infty) = 0$, $\varphi^{(n+1)} \geq 0$ and $\psi^{(n+1)} \geq 0$; this implies that $\varphi^{(n+2)}$ and $\psi^{(n+2)}$ exist for $x \geq x_0$ and that the identity (5) holds for step $n+1$.

Since $\psi^{(n)} \geq 0$, we get from (5)

$$\begin{aligned}
\varphi^{(n+1)}(x) &\leq -\frac{t-1}{t+1}\int_{-l}^{l} \varphi^{(n)}(x-y)N_l(y)dy \\
&\leq \frac{t-1}{t+1}\max\{\varphi^{(n)}(y) \,|\, |y-x| \leq l\}\|N_l\|_{L^1} \\
&\leq \frac{t-1}{t+1}\varphi^{(n)}(x-l)\|N_l\|_{L^1} \quad \text{for } x \geq x_0.
\end{aligned}$$

Choose $x \geq x_0 + l$ to get

$$\begin{aligned}
\varphi^{(n+1)}(x-l) = \int_{x-l}^\infty \varphi^{(n)}(y)dy &= \int_{x-l}^x \varphi^{(n)}(y)dy + \varphi^{(n+1)}(x) \\
&\leq \varphi^{(n)}(x-l)l + \frac{t-1}{t+1}\varphi^{(n)}(x-l)\|N_l\|_{L^1} = k^{-1}\varphi^{(n)}(x-l),
\end{aligned}$$

for $n = 1, 2, 3, \ldots$, where

$$k^{-1} = l + \frac{t-1}{t+1}\|N_l\|_{L^1}.$$

Thus

$$\varphi^{(n)}(x) \le k^{-(n-1)}\varphi^{(1)}(x) \quad \text{for } x \ge x_0, \quad n = 1, 2, 3, \ldots.$$

Now, by integration by parts,

$$\varphi^{(n)}(x) = \frac{1}{(n-1)!}\int_x^\infty (y-x)^{n-1}\varphi(y)dy, \qquad n = 1, 2, 3, \ldots.$$

Therefore, for $x \ge x_0$ and $0 < \lambda < k$,

$$\begin{aligned}\int_x^\infty e^{\lambda y}\varphi(y)dy &= e^{\lambda x}\sum_{n=0}^\infty \frac{\lambda^n}{n!}\int_x^\infty (y-x)^n\varphi(y)dy \\ &= e^{\lambda x}\sum_{n=0}^\infty \lambda^n\varphi^{(n+1)}(x) \\ &\le e^{\lambda x}\sum_{n=0}^\infty \lambda^n k^{-n}\varphi^{(1)}(x) < \infty.\end{aligned}$$

This proves (a).

(b) Let $dF(x) = p(x)dx$, where $p(x) \ge 0$ is a bounded measurable function. Let $0 < \lambda < \sigma$. Choose λ_0 and q such that $0 < \lambda < \lambda_0 < \sigma$ and $1 < q < \lambda_0/\lambda$. Let r be the conjugate exponent of q, i.e., $\frac{1}{q} + \frac{1}{r} = 1$. Then

$$\varphi(x)e^{\lambda x} = f(x)\int_{-\infty}^\infty \varphi(x-y)e^{\lambda(x-y)}e^{\lambda y}p(y)dy = f(x)\{A + B\},$$

where

$$A = \int_{-\infty}^x \varphi(x-y)e^{\lambda(x-y)}e^{\lambda y}p(y)dy, \qquad B = \int_x^\infty \varphi(x-y)e^{\lambda(x-y)}e^{\lambda y}p(y)dy.$$

Now

$$\begin{aligned}0 \le B &\le (\sup_{y\le 0}\varphi(y)e^{\lambda y})\int_x^\infty e^{\lambda y}p(y)dy \le \|\varphi\|_\infty \hat{F}(\lambda), \\ A &= \int_{-\infty}^x \varphi(x-y)e^{\frac{\lambda_0}{q}(x-y)}e^{(\lambda-\frac{\lambda_0}{q})(x-y)}e^{\lambda y}p(y)dy \\ &\le \left(\int_{-\infty}^x \varphi(x-y)^q e^{\lambda_0(x-y)}dy\right)^{\frac{1}{q}}\left(\int_{-\infty}^x e^{r(\lambda-\frac{\lambda_0}{q})(x-y)}e^{r\lambda y}p(y)^r dy\right)^{\frac{1}{r}} \\ &\le \|\varphi\|_\infty^{\frac{q-1}{q}}\left(\int_{-\infty}^\infty \varphi(x)e^{\lambda_0 x}dx\right)^{\frac{1}{q}}\|p\|_\infty^{\frac{r-1}{r}}\hat{F}(r\lambda)^{\frac{1}{r}} < \infty.\end{aligned}$$

Therefore $\varphi(x)e^{\lambda x}$ is uniformly bounded. When $\lambda > \sigma$, if $\varphi(x)e^{\lambda x}$ is bounded, $\varphi(x)e^{\frac{1}{2}(\lambda+\sigma)x}$ is integrable, a contradiction to the definition of σ.

(c) We rewrite the equation as

$$\varphi(x) - f(\infty)\,\varphi * F(x) = (f(x) - f(\infty))\varphi * F(x) \equiv h(x).$$

Let $\Phi(\lambda) = \int_{-\infty}^\infty e^{\lambda x}\varphi(x)dx$ and $H(\lambda) = \int_{-\infty}^\infty e^{\lambda x}h(x)dx$ be the Laplace transforms of φ and h, respectively. By (a), $\Phi(\lambda)$ is finite for $0 < \lambda < \sigma$; hence the Laplace transform of $\varphi * F$ is finite for $0 < \lambda < \sigma$. We have the identity

$$[1 - f(\infty)\hat{F}(\lambda)]\Phi(\lambda) = H(\lambda) \quad \text{for } 0 < \lambda < \sigma. \tag{6}$$

Note that $f(\infty)F_0 < 1$. The equation $f(\infty)\hat{F}(\lambda) = 1$ admits at least one positive root. However, $h(x) \leq 0$ and $h(x) \not\equiv 0$ in $\mathbb{R}$. If $\sigma = \infty$, then $\Phi(\lambda)$ and $H(\lambda)$ exist for all $\lambda > 0$ and $H(\lambda) < 0$. Thus the identity (6) can not hold for all $\lambda > 0$. Therefore $0 < \sigma < \infty$. Then $\Phi(\lambda)$ must be singular at $\lambda = \sigma$, i.e., $\Phi(\sigma) = \infty$ (see Widder [17]). Now

$$\begin{aligned}
&\int_{-\infty}^{\infty} |f(x) - f(\infty)| \varphi * F(x) e^{(\sigma + \frac{1}{2}\epsilon)x} dx \\
&= \int_{-\infty}^{\infty} |f(x) - f(\infty)| \left(\int_{-\infty}^{\infty} \varphi(x-y) e^{(\sigma - \frac{1}{2}\epsilon)(x-y)} e^{(\sigma - \frac{1}{2}\epsilon)y} dF(y) \right) e^{\epsilon x} dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |f(x) - f(\infty)| e^{\epsilon x} \varphi(x-y) e^{(\sigma - \frac{1}{2}\epsilon)(x-y)} dx \right) e^{(\sigma - \frac{1}{2}\epsilon)y} \, dF(y).
\end{aligned}$$

Since $|f(x) - f(\infty)| e^{\epsilon x}$ is bounded and $\varphi(x) e^{(\sigma - \frac{1}{2}\epsilon)(x-y)}$ is integrable,

$$\int_{-\infty}^{\infty} |f(x) - f(\infty)| e^{\epsilon x} \varphi(x-y) e^{(\sigma - \frac{1}{2}\epsilon)(x-y)} dx < \infty$$

uniformly in y. Thus $H(\sigma + \frac{1}{2}\epsilon)$ is finite. The above argument also implies that $H(\lambda)$ is finite for $0 < \lambda \leq \sigma + \frac{\epsilon}{2}$. This proves that $f(\infty)\hat{F}(\sigma) = 1$. σ is obviously the smallest positive root of $f(\infty)\hat{F}(\lambda) = 1$.

It remains to prove that $\varphi(x)$ has the desired asymptotic representation. Since $\hat{F}(\lambda)$ is convex with $\hat{F}(0) = 1$, $\hat{F}(\lambda)$ is decreasing near σ. Hence we can choose λ such that $\sigma < \lambda < \sigma + \frac{\epsilon}{2}$ and $0 < l = \hat{F}(\lambda) f(\infty) < 1$. We rewrite the equation as

$$\varphi = f(\infty)\, \varphi * F + (f(x) - f(\infty)) \varphi * F.$$

Multiply both sides by $e^{\lambda x}$. Then $\varphi_\lambda = l\, \varphi_\lambda * F_\lambda + g$, where

$$\begin{aligned}
&\varphi_\lambda = \varphi(x) e^{\lambda x}, \qquad dF_\lambda(x) = \frac{1}{\hat{F}(\lambda)} e^{\lambda x} dF(x), \\
&g(x) = (f(x) - f(\infty)) e^{\lambda x}\, \varphi * F.
\end{aligned}$$

Since f and φ are uniformly bounded, g decays exponentially near $x = -\infty$. Near $x = \infty$,

$$g(x) = O(e^{-\epsilon x}) e^{\lambda x}\, \varphi * F(x) = O(\varphi_{\lambda - \epsilon} * F_{\lambda - \epsilon}).$$

But $0 < \lambda - \epsilon < \sigma - \frac{\epsilon}{2} < \sigma$, and by (b), $\varphi_{\lambda - \epsilon}$ is bounded. Thus $g(x)$ is bounded and continuous, and decays exponentially at $\pm\infty$. Therefore, there exists a unique bounded continuous function $U(x)$ satisfying $U = l\, U * F_\lambda + g$. Indeed,

$$U = g + \sum_{j=1}^{\infty} l^j g * F_\lambda^{(j)},$$

where $F_\lambda^{(j)} = F_\lambda * \cdots * F_\lambda$ (j times). Note that $U(x) \leq 0$ for all $x \in \mathbb{R}$, since $g \leq 0$ by assumption. Define

$$\psi(x) = \varphi(x) - U(x) e^{-\lambda x} \geq 0.$$

ψ satisfies

$$\psi = f(\infty)\,\psi * F, \quad 0 \le \psi \le C(1 + e^{-\lambda x}), \qquad C \text{ some constant.}$$

This equation has the general solution

$$\psi(x) = \sum_{\mu} P_\mu(x) e^{-\mu x}, \qquad P_\mu \text{ a polynomial,}$$

where μ is a (real or complex) root of $1 = f(\infty)\hat{F}(\mu)$. Since σ is the smallest positive root of $1 = f(\infty)\hat{F}(\mu)$ and $l < 1$, the equation $f(\infty)\hat{F}(\mu) = 1$ has only one real root σ inside the strip $0 \le Re(\mu) \le \lambda$. This σ is a simple root of $f(\infty)\hat{F}(\mu) = 1$. All other roots inside the strip $0 \le Re(\mu) \le \lambda$ are pairwise complex conjugate. Since, for $\mu = \xi + i\eta$ where ξ and η are real,

$$\hat{F}(\xi + i\eta) = \int_{-\infty}^{\infty} e^{i\eta x} e^{\xi x} dF(x) \to 0 \qquad \text{as } |\eta| \to \infty,$$

$\hat{F}(\mu) f(\infty) = 1$ has only a finite number of roots inside the strip $0 \le Re(\mu) \le \lambda$. Also, roots μ with $Re(\mu) > \lambda$ are excluded by the requirement that $0 \le \psi \le C(1 + e^{-\lambda x})$. Thus an application of Theorem 146 in Titchmarsh [14, p. 305] gives

$$\psi(x) = Ae^{-\sigma x} + \sum_{\mu} P_\mu(x) e^{-\mu x},$$

where the sum is over complex numbers μ such that $f(\infty)\hat{F}(\mu) = 1$ and $0 \le Re(\mu) \le \lambda$. But $\psi(x) \ge 0$ for all x, and we must have $\psi(x) = Ae^{-\sigma x}$ with $A \ge 0$. Therefore

$$\varphi(x) = Ae^{-\sigma x} + U(x)e^{-\lambda x} = e^{-\sigma x}[A + U(x)e^{-\epsilon' x}],$$

where $\epsilon' = \lambda - \sigma$. This A can not be zero, otherwise $\varphi_\sigma = \varphi e^{\sigma x}$ satisfies

$$0 \le \varphi_\sigma \le \varphi_\sigma * F_\sigma, \qquad \varphi_\sigma(\pm\infty) = 0.$$

Theorem 4.1 of Essén [3, p. 126] implies that $\varphi_\sigma(x) \equiv 0$ for all x, a contradiction to $\varphi \not\equiv 0$.

The proof of Lemma 4.1 is complete. □

5. The existence of traveling waves. From now on, we shall assume (A1) and (A2)′–(A4)′. The mapping Q is simplified to

$$\tilde{N} = \alpha(N * \mathcal{S})\left(\frac{u^2}{N} * \mathcal{S}\right) + 2\gamma(N * \mathcal{S})\left(\frac{uv}{N} * \mathcal{S}\right) + \gamma(N * \mathcal{S})\left(\frac{v^2}{N} * \mathcal{S}\right),$$
$$\tilde{u} = \alpha(N * \mathcal{S})\left(\frac{u^2}{N} * \mathcal{S}\right) + \gamma(N * \mathcal{S})\left(\frac{uv}{N} * \mathcal{S}\right).$$

It is more convenient to work with the variable $v = N - u$ instead of u. In terms of $[N, v]$ and $k(N) = \frac{\alpha(N)}{\gamma(N)}$, the mapping Q is

$$\tilde{N} = \gamma(N * \mathcal{S})N * \mathcal{S} + (k(N * \mathcal{S}) - 1)\gamma(N * \mathcal{S})\left(\frac{u^2}{N}\right) * \mathcal{S},$$
$$\tilde{v} = \gamma(N * \mathcal{S})v * \mathcal{S}.$$

We look for a traveling wave solution of Q in the form

$$N^{(n)}(x) = N(x - nc), \quad v^{(n)} = v(x - nc), \quad n = 0, 1, 2, \ldots,$$

where c is the wave speed and $[N(x), v(x)] \in \Delta$ is the wave profile. Thus $[N(x), v(x)]$ is such a wave profile iff it satisfies the following system of convolution equations:

$$N = \gamma(N * \mathcal{S}_c)\, N * \mathcal{S}_c + (k(N * \mathcal{S}_c) - 1)\gamma(N * \mathcal{S}_c) \left(\frac{u^2}{N}\right) * \mathcal{S}_c \equiv Q_{1c}[N, v],$$
$$v = \gamma(N * \mathcal{S}_c) v * \mathcal{S}_c \equiv Q_{2c}[N, v],$$

where $\mathcal{S}_c(x) = \mathcal{S}(x + c)$ is the translated probability distribution. We will denote $[Q_{1c}, Q_{2c}]$ by Q_c. Clearly $[N(\pm\infty), v(\pm\infty)]$ must be constant fixed points of Q. In this section, we look for traveling wave solutions which satisfy the boundary conditions

$$[N(-\infty), v(-\infty)] = [N^*, N^*], \qquad [N(\infty), v(\infty)] = [N_1^*, 0]. \tag{7}$$

We shall discuss other types of traveling wave solutions that connect $[N^*, N^*]$ and $[0, 0]$ in §7.

LEMMA 5.1. *Let $[N(x), u(x)]$ be a traveling wave of the system Q with speed c which satisfies the boundary conditions* (7). *Suppose that $\alpha'(N_1^*)N_1^* + 1 \geq 0$.*

(a) *Then $0 < N(x) \leq N^*$ for all $x \in \mathbb{R}$. Thus $\inf_{x\in\mathbb{R}} N(x) > 0$.*

(b) *If $\alpha(N^*)N^* \geq N_1^*$, then $N_1^* \leq N(x) \leq N^*$ for all $x \in \mathbb{R}$.*

(c) *If $|\alpha'(N^*)N^* + \alpha(N^*)| \leq \alpha'(N_1^*)N_1^* + 1$, then $c \geq c_1^*$. Moreover, when $c > c_1^*$, $N(x)$ and $v(x)$ have a representation of the form*

$$N(x) = N_1^* + e^{-\sigma x}\{A + R(x)e^{-\epsilon x}\},$$
$$v(x) = e^{-\sigma x}\{B + V(x)e^{-\epsilon x}\},$$

where $R(x)$ and $V(x)$ are bounded continuous functions, σ is the smallest positive root of the equation $\gamma(N_1^*)e^{-\lambda c}K(\lambda) = 1$, and $0 < \epsilon < \sigma$. A and B are two constants with the relation

$$A = \frac{2(\gamma(N_1^*) - 1)}{\gamma(N_1^*) - (\alpha'(N_1^*)N_1^* + 1)} B.$$

Proof. (a) By (A2)′, we have

$$N(x) \leq \gamma(N * \mathcal{S}_c(x))\, N * \mathcal{S}_c(x) \leq M,$$

where $M = \max_{0\leq N\leq 1} N\gamma(N)$. But $N\gamma(N)$ is increasing over $0 \leq N \leq M$. Therefore $N(x) \leq M_n$ for all $x \in \mathbb{R}$, where M_n is defined by

$$M_{n+1} = \gamma(M_n)M_n, \quad M_0 = M, \quad n = 0, 1, 2, \ldots,$$

Since $\gamma'(N^*)N^* + 1 \geq 0$, M_n decreases to N^* as $n \to \infty$. Thus $N(x) \leq N^*$. Now $N(x)$ satisfies the inequality

$$N(x) \geq \alpha(N * \mathcal{S}_c(x))\, N * \mathcal{S}_c(x) \geq \alpha(N^*)N * \mathcal{S}_c(x).$$

If $N(x_0) = 0$ for some x_0 in $\mathbb{R}$, then $N * \mathcal{S}_c^{(n)}(x_0) = 0$ for all $n = 1, 2, 3, \ldots$. An argument involving the concept of "the point of increase" of a distribution function

(see Diekmann and Kaper [2]) implies that $N(x) = 0$ for all $x \in \mathbb{R}$. This is a contradiction to the boundary conditions of $N(x)$. Therefore $N(x) > 0$ for all $x \in \mathbb{R}$.

(b) Let $N_0 = \inf_{x\in\mathbb{R}} N(x)$. Define $N_1 = \alpha(N_0)N_0$. We *claim* that $N(x) \geq \min\{N_1, N_1^*\}$ for all $x \in \mathbb{R}$. Let $x \in \mathbb{R}$ be a point where $N * \mathcal{S}_c(x) \leq N_1^*$, and we have

$$N(x) \geq \alpha(N * \mathcal{S}_c(x))N * \mathcal{S}_c(x) \geq \alpha(N_0)N_0 = N_1.$$

On the other hand, let $x \in \mathbb{R}$ be a point where $N * \mathcal{S}_c(x) > N_1^*$. Since $\alpha(N^*)N^* \geq N_1^*$ and $N\alpha(N)$ is concave, we have

$$N(x) \geq \alpha(N * \mathcal{S}_c(x))N * \mathcal{S}_c(x) \geq N_1^*.$$

This proves the claim.

From the claim, $N_0 \geq \min\{N_1, N_1^*\}$. If $N_0 < N_1^*$, then $N_0 \geq N_1$. But $\alpha'(N_1^*)N_1^* + 1 \geq 0$ and $N\alpha(N)$ is concave, we must have $N_0 < \alpha(N_0)N_0 = N_1$. This is a contradiction. Therefore $N_0 \geq N_1^*$.

(c) The equation for $v(x)$ is given by

$$v = \gamma(N * \mathcal{S}_c)v * \mathcal{S}_c, \quad v(-\infty) = N^*, \quad v(\infty) = 0.$$

Let $f(x) = \gamma(N * \mathcal{S}_c(x))$. Then $f(\infty) = \gamma(N_1^*) > 1$. By Lemma 4.1 (a) and (b), $v(x)e^{\lambda x}$ is bounded for all $\lambda > 0$ sufficiently small. We *claim* that $(N(x) - N_1^*)e^{\lambda x}$ is also bounded in $\mathbb{R}$. To prove the claim, we write the N equation as

$$\begin{aligned} N - N_1^* = {} & (\alpha(N * \mathcal{S}_c)N * \mathcal{S}_c - N_1^*) + 2(\gamma(N * \mathcal{S}_c) - \alpha(N * \mathcal{S}_c))v * \mathcal{S}_c \\ & + (\alpha(N * \mathcal{S}_c) - \gamma(N * \mathcal{S}_c)) \left(\frac{v^2}{N} * \mathcal{S}_c\right). \end{aligned}$$

By the Mean Value Theorem,

$$\alpha(N * \mathcal{S}_c(x))N * \mathcal{S}_c(x) - N_1^* = T(x)(N * \mathcal{S}_c(x) - N_1^*),$$

where

$$\begin{aligned} T(x) &= \int_0^1 \{\alpha'(N(t) * \mathcal{S}_c)N(t) * \mathcal{S}_c + \alpha(N(t) * \mathcal{S}_c)\}dt, \\ N(t) &= (1-t)N(x) + tN_1^*. \end{aligned}$$

For convenience, we call $N_\lambda(x) = (N(x) - N_1^*)e^{\lambda x}$. We multiply both sides of the N equation by $e^{\lambda x}$. Then

$$N_\lambda(x) = T(x)e^{-\lambda c}K(\lambda)\, N_\lambda * F(x) + g(x), \tag{8}$$

where

$$g(x) = (\gamma(N * \mathcal{S}_c(x)) - \alpha(N * \mathcal{S}_c(x)))e^{-\lambda c}K(\lambda) \left\{2v_\lambda - \frac{v}{N}v_\lambda\right\} * F(x)$$

with

$$v_\lambda(x) = v(x)e^{\lambda x} \quad \text{and} \quad dF(x) = \frac{1}{e^{-\lambda c}K(\lambda)}e^{\lambda x}d\mathcal{S}_c(x).$$

Since $0 \le v \le N$ and v_λ is bounded, $g(x)$ is bounded in $\mathbb{R}$. By (b), $N_1^* \le N(t) \le N^*$. Note that $\alpha'(N)N+\alpha(N)$ is decreasing in N and $|\alpha'(N^*)N^*+\alpha(N^*)| \le \alpha'(N_1^*)N_1^*+1$. Therefore $|T(x)| \le \alpha'(N_1^*)N_1^* + 1$ for all $x \in \mathbb{R}$. Since

$$\lim_{\lambda \to 0}(\alpha'(N_1^*)N_1^* + 1)e^{-\lambda c}K(\lambda) = \alpha'(N_1^*)N_1^* + 1 < 1,$$

we have $(\alpha'(N_1^*)N_1^* + 1)e^{-\lambda c}K(\lambda) < 1$ for all $\lambda > 0$ small enough. Thus, for such λ, we get

$$|T(x)e^{-\lambda c}K(\lambda)| \le (\alpha'(N_1^*)N_1^* + 1)e^{-\lambda c}K(\lambda) < 1$$

for all $x \in \mathbb{R}$. Therefore the equation

$$\varphi(x) = T(x)e^{-\lambda c}K(\lambda)\,\varphi * F(x) + g(x)$$

has a unique solution which is bounded in $\mathbb{R}$. This proves that $N_\lambda(x)$ is bounded in $\mathbb{R}$ for some $\lambda > 0$.

From the claim, we deduce that

$$f(x) = \gamma(N * \mathcal{S}_c(x)) = \gamma(N_1^*) + O(e^{-\lambda x})$$

as $x \to \infty$ for some $\lambda > 0$. Thus Lemma 4.1 (c) implies that the equation $\gamma(N_1^*)e^{-\lambda c}K(\lambda) = 1$ has a smallest positive root σ. Therefore

$$c = \frac{1}{\sigma}\log(\gamma(N_1^*)K(\sigma)) \ge c_1^*.$$

Moreover, when $c > c_1^*$, $v(x)$ has a representation of the form

$$v(x) = e^{-\sigma x}\{B + V(x)e^{-\epsilon x}\}$$

for some $0 < \epsilon < \sigma$, where $B > 0$ and $V(x)$ is a bounded continuous function with $V(x) \le 0$ for all $x \in \mathbb{R}$. We put this form of $v(x)$ into the N equation. By a similar argument as in the proof of the claim, we get the representation of $N(x)$. Indeed, we will get the relation between A and B as follows. Since $A = \lim_{x\to\infty}(N(x) - N_1^*)e^{\sigma x}$, we get

$$A = (\alpha'(N_1^*)N_1^* + 1)e^{-\sigma c}K(\sigma)A + 2(\gamma(N_1^*) - \alpha(N_1^*))e^{-\sigma c}K(\sigma)B$$

by letting $x \to \infty$ in (8) for $\lambda = \sigma$. Thus (c) is proved. □

LEMMA 5.2. *Let* $[N(x), u(x)] \in \Delta$ *have the asymptotic form*

$$N(x) = L + e^{-\sigma x}\{A + e^{-\epsilon x}R(x)\}, \qquad v(x) = e^{-\sigma x}\{B + e^{-\epsilon x}V(x)\}$$

for some $0 < \epsilon < \sigma$. $R(x)$ *and* $V(x)$ *are bounded continuous functions.* L, A, *and* B *are constants. Assume that* $\inf_{x\in\mathbb{R}} N(x) > 0$. *Then*

$$\tilde{N}(x) = Q_{1c}[N, v](x) = \tilde{L} + e^{-\sigma x}\{\tilde{A} + e^{-\epsilon x}\tilde{R}(x)\},$$
$$\tilde{v}(x) = Q_{2c}[N, v](x) = e^{-\sigma x}\{\tilde{B} + e^{-\epsilon x}\tilde{V}(x)\}.$$

$\tilde{L}$, $\tilde{A}$, *and* $\tilde{B}$ *are computed as*

$$\tilde{L} = \alpha(L)L,$$
$$\tilde{A} = e^{-\sigma c}K(\sigma)[A(\alpha'(L)L + \alpha(L)) + 2B(\gamma(L) - \alpha(L))],$$
$$\tilde{B} = e^{-\sigma c}K(\sigma)\gamma(L)B.$$

$\tilde{R}$ and $\tilde{V}(x)$ are also bounded continuous functions.

Proof. We only prove the representation of $\tilde{N}$. The argument for $\tilde{v}$ is similar. We write

$$(Q_{1c}[N,v]-\tilde{L})e^{\sigma x}=(\alpha(N*\mathcal{S}_c)N*\mathcal{S}_c-\alpha(L)L)e^{\sigma x}+E[N,v]e^{\sigma x},$$

where

$$E[N,v]=(\gamma(N*\mathcal{S}_c)-\alpha(N*\mathcal{S}_c))\left\{2v-\frac{v^2}{N}\right\}*\mathcal{S}_c.$$

But

$$\begin{aligned}&(\alpha(N*\mathcal{S}_c(x))N*\mathcal{S}_c(x)-\alpha(L)L)e^{\sigma x}\\&\quad=T(x)(N*\mathcal{S}_c(x)-L)e^{\sigma x}\\&\quad=T(x)[Ae^{-\sigma c}K(\sigma)+e^{-(\sigma+\epsilon)c}K(\sigma+\epsilon)R*F_1(x)e^{-\epsilon x}]\\&\quad=A(\alpha'(L)L+\alpha(L))e^{-\sigma c}K(\sigma)+B_1(x)+C_1(x)e^{-\epsilon x},\end{aligned}$$

where

$$\begin{aligned}T(x)&=\int_0^1\{\alpha'(N(t)*\mathcal{S}_c)N(t)*\mathcal{S}_c+\alpha(N(t)*\mathcal{S}_c)\}dt,\\N(t)&=(1-t)N(x)+tL,\\B_1(x)&=[T(x)-(\alpha'(L)L+\alpha(L))]e^{-\sigma c}K(\sigma)A,\\C_1(x)&=T(x)e^{-(\sigma+\epsilon)c}K(\sigma+\epsilon)R*F_1(x),\end{aligned}$$

and

$$dF_1(x)=(e^{-(\sigma+\epsilon)c}K(\sigma+\epsilon))^{-1}e^{(\sigma+\epsilon)x}d\mathcal{S}_c(x).$$

Since $N(t)=L+O(e^{-\sigma x})$ as $x\to\infty$, we have

$$T(x)=(\alpha'(L)L+\alpha(L))+O(e^{-\sigma x})$$

as $x\to\infty$. Thus $B_1(x)$ is a bounded continuous function with $B_1(x)=O(e^{-\sigma x})$ as $x\to\infty$. Since $T(x)$ is bounded, $C_1(x)$ is also a bounded continuous function. With a similar computation, we can write

$$E[N,v](x)e^{\sigma x}=2(\gamma(L)-\alpha(L))K(\sigma)e^{-\sigma c}B+B_2(x)+C_2(x)e^{-\epsilon x},$$

where

$$\begin{aligned}B_2(x)&=2(\gamma(N*\mathcal{S}_c(x))-\alpha(N*\mathcal{S}_c(x))-\gamma(L)+\alpha(L))K(\sigma)e^{-\sigma c}B\\&\quad-(\gamma(N*\mathcal{S}_c(x))-\alpha(N*\mathcal{S}_c(x)))\left(\frac{v^2}{N}\right)*\mathcal{S}_c(x)e^{\sigma x},\\C_2(x)&=2(\gamma(N*\mathcal{S}_c(x))-\alpha(N*\mathcal{S}_c(x)))e^{-(\sigma+\epsilon)c}K(\sigma+\epsilon)V*F_1(x).\end{aligned}$$

$C_2(x)$ is obviously bounded. Now

$$\gamma(N*\mathcal{S}_c(x))-\alpha(N*\mathcal{S}_c(x))=\gamma(L)-\alpha(L)+O(e^{-\sigma x}),$$

$v^2(x) = O(e^{-2\sigma x})$ as $x \to \infty$, and $\inf_{x\in\mathbb{R}} N(x) > 0$. We get

$$B_2(x) = O(e^{-\sigma x}) \quad \text{as } x \to \infty.$$

Therefore

$$\begin{aligned}[Q_{1c}[N,v](x) - \tilde{L}]e^{\sigma x} &= A(\alpha'(L)L + \alpha(L))e^{-\sigma c}K(\sigma) \\ &\quad +2(\gamma(L) - \alpha(L))K(\sigma)e^{-\sigma c}B \\ &\quad +(B_1(x) + B_2(x)) + (C_1(x) + C_2(x))e^{-\epsilon x} \\ &= \tilde{A} + e^{-\epsilon x}\tilde{R}(x),\end{aligned}$$

where

$$\tilde{R}(x) = (B_1(x) + B_2(x))e^{\epsilon x} + (C_1(x) + C_2(x)).$$

However, $0 < \epsilon < \sigma$. Hence

$$(B_1(x) + B_2(x))e^{\epsilon x} = O(e^{-(\sigma-\epsilon)x})$$

as $x \to \infty$. Therefore $\tilde{R}(x)$ is bounded. The proof is complete. □

We now establish a key lemma on the Lipschitz estimate of Q.

LEMMA 5.3. *Assume* (A4)′. *Then there exist* $0 < k_0 < 1$ *and* $\eta_0 > 0$ *with the following property. If* $k_0 \le k(N) < 1$ *for all* $0 \le N \le N^*$ *and* $\max_{0\le N\le N^*} |k'(N)| \le \eta_0$, *then for any* $[N_i(x), v_i(x)] \in \Delta$ *with* $N_1^* \le N_i(x) \le N^*$ $(i = 1, 2)$, *we have the following inequalities:*

(a) $|\tilde{v}_1(x) - \tilde{v}_2(x)| \le \gamma(N_1^*) \max\{|u_1 - u_2| * \mathcal{S}_c(x),\ |v_1 - v_2| * \mathcal{S}_c(x)\}$,

(b) $|\tilde{u}_1(x) - \tilde{u}_2(x)| \le \gamma(N_1^*)\ \max\{|N_1 - N_2|, |u_1 - u_2|, |v_1 - v_2|\} * \mathcal{S}_c(x)$

for all x *in* $\mathbb{R}$, *where*

$$\tilde{N}_i = Q_{1c}[N_i, v_i], \quad \tilde{v}_i = Q_{2c}[N_i, v_i], \quad \tilde{u}_i = \tilde{N}_i - \tilde{v}_i$$

for $i = 1, 2$.

Proof. We prove the lemma at each point x in $\mathbb{R}$. Without loss of generality, we may assume that $N_1 * \mathcal{S}_c(x) \ge N_2 * \mathcal{S}_c(x)$ at x. The argument for the case where $N_1 * \mathcal{S}_c(x) \le N_2 * \mathcal{S}_c(x)$ is the same. We divide the proof into several steps.

Step 1. We prove the inequality (a).

By the definition of $\tilde{v}_i$ $(i = 1, 2)$, we have

$$\tilde{v}_1 - \tilde{v}_2 = \gamma(N_1 * \mathcal{S}_c)(v_1 - v_2) * \mathcal{S}_c + (\gamma(N_1 * \mathcal{S}_c) - \gamma(N_2 * \mathcal{S}_c))v_2 * \mathcal{S}_c.$$

Since $N_1 * \mathcal{S}_c(x) \ge N_2 * \mathcal{S}_c(x)$ and γ is decreasing, we get $(\gamma(N_1 * \mathcal{S}_c(x)) - \gamma(N_2 * \mathcal{S}_c(x)))v_2 * \mathcal{S}_c(x) \le 0$. Therefore

$$\tilde{v}_1(x) - \tilde{v}_2(x) \le \gamma(N_1^*)|v_1 - v_2| * \mathcal{S}_c(x).$$

Now we obtain a lower bound for $\tilde{v}_1(x) - \tilde{v}_2(x)$:

$$\begin{aligned}\tilde{v}_1 - \tilde{v}_2 &= \gamma(N_1 * \mathcal{S}_c)(N_1 - u_1) * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)(N_2 - u_2) * \mathcal{S}_c \\ &= (\gamma(N_1 * \mathcal{S}_c)N_1 * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)N_2 * \mathcal{S}_c) \\ &\quad +(\gamma(N_2 * \mathcal{S}_c) - \gamma(N_1 * \mathcal{S}_c))u_1 * \mathcal{S}_c + \gamma(N_2 * \mathcal{S}_c)(u_2 - u_1) * \mathcal{S}_c.\end{aligned}$$

Since $N\gamma(N)$ is increasing and $\gamma(N)$ is decreasing in $N_1^* \le N \le N^*$, we have $\gamma(N_1 * \mathcal{S}_c(x))N_1 * \mathcal{S}_c(x) - \gamma(N_2 * \mathcal{S}_c(x))N_2 * \mathcal{S}_c(x) \ge 0$ and $(\gamma(N_2 * \mathcal{S}_c(x)) - \gamma(N_1 * \mathcal{S}_c(x)))u_1 * \mathcal{S}_c(x) \ge 0$. Then

$$\tilde{v}_1(x) - \tilde{v}_2(x) \ge \gamma(N_2 * \mathcal{S}_c(x))(u_2 - u_1) * \mathcal{S}_c(x) \ge -\gamma(N_1^*)|u_1 - u_2| * \mathcal{S}_c(x).$$

The proof of Step 1 is complete. Note that the inequality (a) for $\tilde{v}$ is independent of k_0 and η_0.

For convenience, we define $\kappa = \min_{0\le N\le N^*} k(N)$ and $\eta = \max_{0\le N\le N^*} |k'(N)|$.

Step 2. There exist $0 < k_1 < 1$ and $0 < \eta_2$ such that whenever $k_1 \le \kappa$ and $0 \le \eta \le \eta_2$, the following inequality holds:

$$\gamma(N_1 * \mathcal{S}_c(x))\tilde{H}(x,\cdot) * \mathcal{S}_c(x) \le \tilde{u}_1(x) - \tilde{u}_2(x) \le \gamma(N_1 * \mathcal{S}_c(x))H(x,\cdot) * \mathcal{S}_c(x)$$

for all x in $\mathbb{R}$, where

$$\begin{aligned} H(x,y) &= (u_1(y) - u_2(y))\left[1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\frac{u_1(y) + u_2(y)}{N_1(y)}\right] \\ &\quad + (k(N_1 * \mathcal{S}_c(x)) - 1)\frac{u_2(y)^2}{N_1(y)N_2(y)}(N_2(y) - N_1(y)), \\ \tilde{H}(x,y) &= (v_2(y) - v_1(y))\left[1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\left(2 - \frac{v_1(y) + v_2(y)}{N_1(y)}\right)\right] \\ &\quad + (k(N_1 * \mathcal{S}_c(x)) - 1)\frac{v_2(y)^2}{N_1(y)N_2(y)}(N_2(y) - N_1(y)) \end{aligned}$$

and $H(x,\cdot) * \mathcal{S}_c(x)$ is defined by

$$H(x,\cdot) * \mathcal{S}_c(x) = \int_{-\infty}^{\infty} H(x, x-y)d\mathcal{S}_c(y).$$

$\tilde{H}(x,\cdot) * \mathcal{S}_c(x)$ is similarly defined.

To find the upper bound, we write

$$\begin{aligned} \tilde{u}_1 - \tilde{u}_2 &= (\gamma(N_1 * \mathcal{S}_c)u_1 * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)u_2 * \mathcal{S}_c) \\ &\quad + \left[(k(N_1 * \mathcal{S}_c) - 1)\gamma(N_1 * \mathcal{S}_c)\frac{u_1^2}{N_1} * \mathcal{S}_c\right. \\ &\qquad \left. -(k(N_2 * \mathcal{S}_c) - 1)\gamma(N_2 * \mathcal{S}_c)\frac{u_2^2}{N_2} * \mathcal{S}_c\right] \\ &= \gamma(N_1 * \mathcal{S}_c)(u_1 - u_2) * \mathcal{S}_c + u_2 * \mathcal{S}_c(\gamma(N_1 * \mathcal{S}_c) - \gamma(N_2 * \mathcal{S}_c)) \\ &\quad + (k(N_1 * \mathcal{S}_c) - 1)\left[\gamma(N_1 * \mathcal{S}_c)\left(\frac{u_1^2}{N_1} - \frac{u_2^2}{N_2}\right) * \mathcal{S}_c\right. \\ &\qquad \left. + (\gamma(N_1 * \mathcal{S}_c) - \gamma(N_2 * \mathcal{S}_c))\frac{u_2^2}{N_2} * \mathcal{S}_c\right] \\ &\quad + (k(N_1 * \mathcal{S}_c) - k(N_2 * \mathcal{S}_c))\gamma(N_2 * \mathcal{S}_c)\frac{u_2^2}{N_2} * \mathcal{S}_c \\ &= \gamma(N_1 * \mathcal{S}_c)\left\{(u_1 - u_2) * \mathcal{S}_c + (k(N_1 * \mathcal{S}_c) - 1)\left[\frac{u_1 + u_2}{N_1}(u_1 - u_2)\right.\right. \\ &\qquad \left.\left. + \frac{u_2^2}{N_1N_2}(N_2 - N_1)\right] * \mathcal{S}_c\right\} \end{aligned}$$

$$+\left[u_2 * \mathcal{S}_c + (k(N_1 * \mathcal{S}_c) - 1)\frac{u_2^2}{N_2} * \mathcal{S}_c\right](\gamma(N_1 * \mathcal{S}_c) - \gamma(N_2 * \mathcal{S}_c))$$
$$+(k(N_1 * \mathcal{S}_c) - k(N_2 * \mathcal{S}_c))\gamma(N_2 * \mathcal{S}_c)\frac{u_2^2}{N_2} * \mathcal{S}_c$$
$$= \gamma(N_1 * \mathcal{S}_c)H(x,\cdot) * \mathcal{S}_c + G,$$

where

$$H(x,y) = (u_1(y) - u_2(y))\left[1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\frac{u_1(y) + u_2(y)}{N_1(y)}\right]$$
$$+(k(N_1 * \mathcal{S}_c(x)) - 1)\frac{u_2(y)^2}{N_1(y)N_2(y)}(N_2(y) - N_1(y)),$$
$$G = \left[u_2 * \mathcal{S}_c + (k(N_1 * \mathcal{S}_c) - 1)\frac{u_2^2}{N_2} * \mathcal{S}_c\right](\gamma(N_1 * \mathcal{S}_c) - \gamma(N_2 * \mathcal{S}_c))$$
$$+(k(N_1 * \mathcal{S}_c) - k(N_2 * \mathcal{S}_c))\gamma(N_2 * \mathcal{S}_c)\frac{u_2^2}{N_2} * \mathcal{S}_c.$$

We *claim* that there exists $\eta_1 > 0$ such that $G \leq 0$ whenever $0 \leq \eta \leq \eta_1$. We apply the Mean Value Theorem to rewrite G as

$$G(x) = (N_1 - N_2) * \mathcal{S}_c(x)\{\gamma'(\bullet)[u_2 * \mathcal{S}_c(x)$$
$$+(k(N_1 * \mathcal{S}_c(x)) - 1)\frac{u_2^2}{N_2} * \mathcal{S}_c(x)]$$
$$+k'(\bullet)\gamma(N_2 * \mathcal{S}_c(x))\frac{u_2^2}{N_2} * \mathcal{S}_c(x)\}$$
$$\leq (N_1 - N_2) * \mathcal{S}_c(x)\left\{\gamma'(\bullet)\left(u_2 - \frac{u_2^2}{N_2}\right) * \mathcal{S}_c(x)\right.$$
$$\left.+\frac{u_2^2}{N_2} * \mathcal{S}_c(x)[\gamma'(\bullet)k(N_1 * \mathcal{S}_c(x)) + \gamma(N_1^*)\eta]\right\}.$$

(Recall that $(N_1 - N_2) * \mathcal{S}_c(x) \geq 0$.) Since $\gamma(N)$ is decreasing in $0 \leq N < 1$, we have

$$\gamma'(\bullet)\left(u_2 - \frac{u_2^2}{N_2}\right) * \mathcal{S}_c(x) \leq 0.$$

When $\kappa \geq \frac{1}{2}$, we choose η_1 such that

$$0 < \eta_1 < \frac{1}{2\gamma(0)}\left(\min_{0 \leq N \leq N^*} |\gamma'(N)|\right).$$

Then, for $0 \leq \eta \leq \eta_1$, we have

$$\gamma'(\bullet)k(N_1 * \mathcal{S}_c(x)) + \gamma(N_1^*)\eta \leq 0.$$

Therefore, $G \leq 0$. Thus

$$\tilde{u}_1(x) - \tilde{u}_2(x) \leq \gamma(N_1 * \mathcal{S}_c(x))H(x,\cdot) * \mathcal{S}_c(x).$$

To find the lower bound, we use $u = N - v$. Then

$\tilde{u}_1 - \tilde{u}_2$

$$
\begin{aligned}
&= \gamma(N_1 * \mathcal{S}_c)(N_1 - v_1) * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)(N_2 - v_2) * \mathcal{S}_c \\
&\quad + (k(N_1 * \mathcal{S}_c) - 1)\left[\gamma(N_1 * \mathcal{S}_c)\frac{(N_1 - v_1)^2}{N_1} * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)\frac{(N_2 - v_2)^2}{N_2} * \mathcal{S}_c\right] \\
&\quad + (k(N_1 * \mathcal{S}_c) - k(N_2 * \mathcal{S}_c))\gamma(N_2 * \mathcal{S}_c)\frac{(N_2 - v_2)^2}{N_2} * \mathcal{S}_c \\
&= k(N_1 * \mathcal{S}_c)(\gamma(N_1 * \mathcal{S}_c)N_1 * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)N_2 * \mathcal{S}_c) \\
&\quad + [1 + 2(k(N_1 * \mathcal{S}_c) - 1)](\gamma(N_2 * \mathcal{S}_c)v_2 * \mathcal{S}_c - \gamma(N_1 * \mathcal{S}_c)v_1 * \mathcal{S}_c) \\
&\quad + (k(N_1 * \mathcal{S}_c) - 1)\left(\gamma(N_1 * \mathcal{S}_c)\frac{v_1^2}{N_1} * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)\frac{v_2^2}{N_2} * \mathcal{S}_c\right) \\
&\quad + (k(N_1 * \mathcal{S}_c) - k(N_2 * \mathcal{S}_c))\gamma(N_2 * \mathcal{S}_c)\frac{u_2^2}{N_2} * \mathcal{S}_c \\
&= k(N_1 * \mathcal{S}_c)(\gamma(N_1 * \mathcal{S}_c)N_1 * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)N_2 * \mathcal{S}_c)) \\
&\quad + (k(N_1 * \mathcal{S}_c) - k(N_2 * \mathcal{S}_c))\gamma(N_2 * \mathcal{S}_c)\frac{u_2^2}{N_2} * \mathcal{S}_c \\
&\quad + [(1 + 2(k(N_1 * \mathcal{S}_c) - 1))v_2 * \mathcal{S}_c \\
&\qquad + (1 - k(N_1 * \mathcal{S}_c))\frac{v_2^2}{N_2} * \mathcal{S}_c](\gamma(N_2 * \mathcal{S}_c) - \gamma(N_1 * \mathcal{S}_c)) \\
&\quad + \gamma(N_1 * \mathcal{S}_c)[1 + 2(k(N_1 * \mathcal{S}_c) - 1)](v_2 - v_1) * \mathcal{S}_c \\
&\quad + \gamma(N_1 * \mathcal{S}_c)(k(N_1 * \mathcal{S}_c) - 1)\left[(v_1 - v_2)\frac{v_1 + v_2}{N_1} + \frac{v_2^2}{N_1 N_2}(N_2 - N_1)\right] * \mathcal{S}_c \\
&= \tilde{G} + \gamma(N_1 * \mathcal{S}_c)\tilde{H}(x, \cdot) * \mathcal{S}_c,
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{G} &= k(N_1 * \mathcal{S}_c)(\gamma(N_1 * \mathcal{S}_c)N_1 * \mathcal{S}_c - \gamma(N_2 * \mathcal{S}_c)N_2 * \mathcal{S}_c) \\
&\quad + (k(N_1 * \mathcal{S}_c) - k(N_2 * \mathcal{S}_c))\gamma(N_2 * \mathcal{S}_c)\frac{u_2^2}{N_2} * \mathcal{S}_c \\
&\quad + [(2k(N_1 * \mathcal{S}_c) - 1)v_2 * \mathcal{S}_c \\
&\qquad + (1 - k(N_1 * \mathcal{S}_c))\frac{v_2^2}{N_2} * \mathcal{S}_c](\gamma(N_2 * \mathcal{S}_c) - \gamma(N_1 * \mathcal{S}_c)), \\
\tilde{H}(x, y) &= (v_2(y) - v_1(y))\left[1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\left(2 - \frac{v_1(y) + v_2(y)}{N_1(y)}\right)\right] \\
&\quad + (k(N_1 * \mathcal{S}_c(x)) - 1)\frac{v_2(y)^2}{N_1(y)N_2(y)}(N_2(y) - N_1(y)).
\end{aligned}
$$

We *claim* that there exist $0 < k_1 < 1$ and $0 < \eta_2 \leq \eta_1$ such that $\tilde{G} \geq 0$ whenever $k_1 \leq \kappa$ and $0 \leq \eta \leq \eta_2$. Since $\gamma(N)$ is decreasing and we are assuming that $N_1 * \mathcal{S}_c(x) \geq N_2 * \mathcal{S}_c(x)$, we have $\gamma(N_2 * \mathcal{S}_c(x)) - \gamma(N_1 * \mathcal{S}_c(x)) \geq 0$. If $\frac{1}{2} \leq \kappa$, then

$$
[2k(N_1 * \mathcal{S}_c(x)) - 1]v_2 * \mathcal{S}_c(x) + (1 - k(N_1 * \mathcal{S}_c(x)))\frac{v_2^2}{N_2} * \mathcal{S}_c(x) \geq 0.
$$

Now, by the Mean Value Theorem, we get

$$
\begin{aligned}
\tilde{G}(x) &\geq k(N_1 * \mathcal{S}_c(x))(\gamma'(\xi)\xi + \gamma(\xi))(N_1 - N_2) * \mathcal{S}_c(x) \\
&\quad + \gamma(N_2 * \mathcal{S}_c(x))\frac{u_2^2}{N_2} * \mathcal{S}_c(x)k'(\bullet)(N_1 - N_2) * \mathcal{S}_c(x)
\end{aligned}
$$

for some ξ with $N_1 * \mathcal{S}_c(x) \geq \xi \geq N_2 * \mathcal{S}_c(x)$. Then, when $\kappa \geq \frac{1}{2}$,

$$\begin{aligned}\tilde{G}(x) &\geq \left[\kappa \min_{N_1^* \leq N \leq N^*}(\gamma'(N)N + \gamma(N)) - N^*\eta\right](N_1 - N_2) * \mathcal{S}_c(x) \\ &\geq \left[\frac{1}{2}(\gamma'(N^*)N^* + 1) - N^*\eta\right](N_1 - N_2) * \mathcal{S}_c(x).\end{aligned}$$

By assumption, $\gamma'(N^*)N^* + 1 > 0$. We define

$$\eta_2 = \min\left\{\eta_1, \frac{1}{2N^*}(\gamma'(N^*)N^* + 1)\right\}.$$

Then, for $\kappa \geq \frac{1}{2}$ and $0 \leq \eta \leq \eta_2$, we have $\tilde{G}(x) \geq 0$. This proves the claim and Step 2 by letting $k_1 = \frac{1}{2}$.

Step 3. There exists $1 > k_2 \geq k_1$ such that whenever $k_2 \leq \kappa$, H satisfies

$$H(x, y) \leq \max\{|N_1(y) - N_2(y)|, |u_1(y) - u_2(y)|\}$$

for all $x, y \in \mathbb{R}$.

We note that

$$1 + (\kappa - 1)\frac{2N^*}{N_1^*} \leq 1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\frac{u_1(y) + u_2(y)}{N_1(y)} \leq 1.$$

When $\kappa \to 1$, $1 + (\kappa - 1)\frac{2N^*}{N_1^*} \to 1$. Thus we can choose $1 > k_2 \geq k_1$ such that

$$0 \leq 1 + (\kappa - 1)\frac{2N^*}{N_1^*} \qquad \text{for } k_2 \leq \kappa.$$

Then, for $\kappa \geq k_2$,

$$\begin{aligned}H(x, y) &\leq |u_1(y) - u_2(y)|\left[1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\frac{u_1(y) + u_2(y)}{N_1(y)}\right] \\ &\quad + |N_1(y) - N_2(y)|(1 - k(N_1 * \mathcal{S}_c(x)))\frac{u_2(y)^2}{N_1(y)N_2(y)} \\ &\leq \max\{|N_1(y) - N_2(y)|, |u_1(y) - u_2(y)|\} \\ &\quad \times \left[1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\left(\frac{u_1(y)}{N_1(y)} + \frac{u_2(y)v_2(y)}{N_1(y)N_2(y)}\right)\right] \\ &\leq \max\{|N_1(y) - N_2(y)|, |u_1(y) - N_2(y)|\}.\end{aligned}$$

Step 4. There exists $1 > k_4 \geq k_2$ such that whenever $k_4 \leq \kappa$, $\tilde{H}$ satisfies

$$\tilde{H}(x, y) \geq -\max\{|N_1(y) - N_2(y)|, |u_1(y) - u_2(y)|, |v_1(y) - v_2(y)|\}$$

for all $x, y \in \mathbb{R}$.

We consider three cases.

Case a. $N_1(y) \geq N_2(y)$.

We have

$$\frac{v_2(y)}{N_1(y)} \leq \frac{v_2(y)}{N_2(y)}, \quad \text{i.e. } 1 - \frac{v_2(y)}{N_1(y)} \geq \frac{u_2(y)}{N_2(y)} \geq 0.$$

Thus

$$1+(k(N_1*\mathcal{S}_c(x))-1)\left(2-\frac{v_1(y)+v_2(y)}{N_1(y)}\right)$$
$$=1+(k(N_1*\mathcal{S}_c)-1)\left[\left(1-\frac{v_1(y)}{N_1(y)}\right)+\left(1-\frac{v_2(y)}{N_1(y)}\right)\right]\le 1.$$

On the other hand, by $\kappa\ge\frac{1}{2}$,

$$1+(k(N_1*\mathcal{S}_c(x))-1)\left(2-\frac{v_1(y)+v_2(y)}{N_1(y)}\right)$$
$$\ge 1+2(\kappa-1)\left(1-\frac{N_1^*}{N^*}\right)\ge 1-\left(1-\frac{N_1^*}{N^*}\right)\ge 0.$$

Since $N_1(y)\ge N_2(y)$,

$$(k(N_1*\mathcal{S}_c(x))-1)\frac{v_2(y)^2}{N_1(y)N_2(y)}(N_2(y)-N_1(y))\ge 0.$$

Therefore

$$\tilde{H}(x,y)\ge(v_2(y)-v_1(y))\left[1+(k(N_1*\mathcal{S}_c(x))-1)\left(2-\frac{v_1(y)+v_2(y)}{N_1(y)}\right)\right]$$
$$\ge-|v_1(y)-v_2(y)|.$$

Case b. $N_1(y)<N_2(y)$ and $v_1(y)\ge v_2(y)$.
Then $u_1(y)\le u_2(y)$. We write $\tilde{H}$ as

$$\tilde{H}(x,y)=(v_2(y)-v_1(y))\left[1+(k(N_1*\mathcal{S}_c(x))-1)\left(2-\frac{v_1(y)+v_2(y)}{N_1(y)}\right)\right]$$
$$+(k(N_1*\mathcal{S}_c(x))-1)\frac{v_2(y)^2}{N_1(y)N_2(y)}(v_2(y)+u_2(y)-v_1(y)-u_1(y))$$
$$=(v_2(y)-v_1(y))\left\{1+(k(N_1*\mathcal{S}_c(x))-1)\left[\frac{u_1(y)}{N_1(y)}\right.\right.$$
$$\left.\left.+\left(1-\frac{v_2(y)}{N_1(y)}\right)+\frac{v_2(y)^2}{N_1(y)N_2(y)}\right]\right\}$$
$$+(k(N_1*\mathcal{S}_c(x))-1)\frac{v_2(y)^2}{N_1(y)N_2(y)}(u_2(y)-u_1(y)).$$

Similar to the argument in Step 3, we can choose $1>k_3\ge k_2$ such that

$$1+(k(N_1*\mathcal{S}_c(x))-1)\left[\frac{u_1(y)}{N_1(y)}+\left(1-\frac{v_2(y)}{N_1(y)}\right)+\frac{v_2(y)^2}{N_1(y)N_2(y)}\right]\ge 0$$

for $k_3\le\kappa$. By $v_1(y)\ge v_2(y)$, we have

$$0\le 1-\frac{v_1(y)}{N_1(y)}\le 1-\frac{v_2(y)}{N_1(y)}.$$

Hence

$$\tilde{H}(x,y)\ge-\max\{|v_1(y)-v_2(y)|,(u_2(y)-u_1(y))\}$$
$$\times\left\{1+(k(N_1*\mathcal{S}_c(x))-1)\left[\frac{u_1(y)}{N_1(y)}+\left(1-\frac{v_2(y)}{N_1(y)}\right)\right]\right\}$$
$$\ge-\max\{|v_1(y)-v_2(y)|,(u_2(y)-u_1(y))\}.$$

Case c. $N_1(y) < N_2(y)$ and $v_1(y) < v_2(y)$.

By the choice of k_3 in Case b, for $\kappa \geq k_3$, we have

$$1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\left(2 - \frac{v_1(y) + v_2(y)}{N_1(y)}\right)$$
$$\geq 1 + (k(N_1 * \mathcal{S}_c(x)) - 1)\left[\frac{u_1(y)}{N_1(y)} + \left(1 - \frac{v_2(y)}{N_1(y)}\right) + \frac{v_2(y)^2}{N_1(y)N_2(y)}\right] \geq 0.$$

Thus

$$\tilde{H}(x,y) \geq (k(N_1 * \mathcal{S}_c(x)) - 1)\frac{v_2(y)^2}{N_1(y)N_2(y)}(N_2(y) - N_1(y)).$$

Now

$$0 \leq (1 - k(N_1 * \mathcal{S}_c(x)))\frac{v_2(y)^2}{N_1(y)N_2(y)} \leq (1-\kappa)\frac{N^*}{N_1^*} \to 0$$

as $\kappa \to 1$. We can choose $1 > k_4 \geq k_3$ such that $(1-\kappa)\frac{N^*}{N_1^*} < 1$ for $\kappa \geq k_4$. Therefore

$$\tilde{H}(x,y) \geq N_1(y) - N_2(y)$$

for $\kappa \geq k_4$. This proves Step 4.

We define $k_0 = k_4$ and $\eta_0 = \eta_2$. By Steps 2, 3, and 4, we get

$$|\tilde{u}_1(x) - \tilde{u}_2(x)| \leq \gamma(N_1^*) \max\{|N_1 - N_2|, |u_1 - u_2|, |v_1 - v_2|\} * \mathcal{S}_c(x)$$

for all $x \in \mathbb{R}$, whenever $k_0 \leq k(N) < 1$ for all $0 \leq N \leq N^*$ and $0 \leq \eta \leq \eta_0$. □

We are now ready to prove Theorem 2.2.

Proof of Theorem 2.2. We choose k_0 and η_0 such that Lemma 5.3 holds. We can also adjust k_0 and η_0 such that $\alpha'(N^*)N^* + \alpha(N^*) \geq 0$ and $\alpha(N^*)N^* \geq N_1^*$ for $k_0 \leq \kappa = \min_{0 \leq N \leq N^*} k(N)$ and $0 \leq \eta \leq \eta_0$. Then Lemma 5.1 can be applied. The necessity that $c \geq c_1^*$ follows. The desired traveling wave has the exponential decay form at $x = \infty$ when $c > c_1^*$.

We prove the existence for $c > c_1^*$ first. Define σ to be the smallest positive root of the equation

$$\gamma(N_1^*)e^{-\lambda c}K(\lambda) = 1.$$

σ exists due to $c > c_1^*$. Let $0 < \epsilon < \sigma$ be chosen such that

$$0 < \gamma(N_1^*)e^{-(\sigma+\epsilon)c}K(\sigma+\epsilon) < 1.$$

Let $\alpha(N)$ and $\gamma(N)$ be given such that κ and η satisfy $k_0 \leq \kappa$ and $0 \leq \eta \leq \eta_0$. For $B > 0$ fixed, we define the space

$$\begin{aligned}\mathcal{B} = \{[R,V] |\ & R(x) \text{ and } V(x) \text{ are bounded continuous functions} \\ & \text{in } \mathbb{R}.\ N(x) = N_1^* + e^{-\sigma x}[A + e^{-\epsilon x}R(x)] \text{ and} \\ & v(x) = e^{-\sigma x}[B + e^{-\epsilon x}V(x)] \text{ satisfy } 0 \leq v(x) \leq N(x) \\ & \text{and } N_1^* \leq N(x) \leq N^* \text{ for all } x \in \mathbb{R}, \}\end{aligned} \tag{9}$$

where

$$A = \frac{2(\gamma(N_1^*) - 1)}{\gamma(N_1^*) - (\alpha'(N_1^*)N_1^* + 1)}B.$$

Clearly $\mathcal{B} \neq \emptyset$. For $[R, V] \in \mathcal{B}$, define

$$u(x) \equiv N(x) - v(x) = N_1^* + e^{-\sigma x}[(A - B) + e^{-\epsilon x}(R(x) - V(x))].$$

We call $U(x) = R(x) - V(x)$. Then we define a metric on $\mathcal{B}$ by

$$\boldsymbol{d}([R_1, V_1], [R_2, V_2]) = \max\{\|R_1 - R_2\|_\infty, \|V_1 - V_2\|_\infty, \|U_1 - U_2\|_\infty\}. \tag{10}$$

$(\mathcal{B}, \boldsymbol{d})$ is then a complete metric space. We define a mapping Γ on $\mathcal{B}$ by

$$\Gamma([R, V]) = [\tilde{R}, \tilde{V}], \qquad [R, V] \in \mathcal{B}, \tag{11}$$

where

$$\begin{aligned} \tilde{R} &= [(Q_{1c}[N, v] - N_1^*)e^{\sigma x} - A]e^{\epsilon x}, \\ \tilde{V} &= [Q_{2c}[N, v]e^{\sigma x} - B]e^{\epsilon x}. \end{aligned}$$

First, we *claim* that $\Gamma : \mathcal{B} \to \mathcal{B}$. By Lemma 5.2, $\tilde{R}$ and $\tilde{V}$ are bounded continuous functions. Now

$$\begin{aligned} [\tilde{R}(x)e^{-\epsilon x} + A]e^{-\sigma x} + N_1^* &= Q_{1c}[N, v](x) \equiv \tilde{N}(x), \\ [\tilde{V}(x)e^{-\epsilon x} + 1]e^{-\sigma x} &= Q_{2c}[N, v](x) \equiv \tilde{v}(x). \end{aligned}$$

From $0 \leq v(x) \leq N(x)$, we get $0 \leq \tilde{v}(x) \leq \tilde{N}(x)$. We need to show that $N_1^* \leq \tilde{N}(x) \leq N^*$ for all $x \in \mathbb{R}$. By the definition of $Q_{1c}[N, v]$, we get

$$\alpha(N * \mathcal{S}_c)N * \mathcal{S}_c \leq \tilde{N} \leq \gamma(N * \mathcal{S}_c)N * \mathcal{S}_c.$$

Because $\gamma'(N^*)N^* + 1 > 0$ and $\alpha(N^*)N^* \geq N_1^*$, we have (from $N_1^* \leq N(x) \leq N^*$)

$$N_1^* \leq \alpha(N * \mathcal{S}_c)N * \mathcal{S}_c \leq \tilde{N}(x) \leq \gamma(N * \mathcal{S}_c)N * \mathcal{S}_c \leq N^*.$$

This proves the claim.

Now we want to prove that Γ is a contraction mapping when k_0 and η_0 are suitably adjusted. Given $[R_i(x), V_i(x)] \in \mathcal{B}$ $(i = 1, 2)$, we let

$$\begin{aligned} N_i(x) &= N_1^* + e^{-\sigma x}[A + e^{-\epsilon x}R_i(x)], \\ v_i(x) &= e^{-\sigma x}[B + e^{-\epsilon x}V_i(x)] \end{aligned}$$

for $i = 1, 2$. We define

$$\tilde{N}_i = Q_{1c}[N_i, v_i], \quad \tilde{v}_i = Q_{2c}[N_i, v_i] \quad \text{for } i = 1, 2.$$

We estimate $\tilde{N}_1(x) - \tilde{N}_2(x)$ as follows.

$$\begin{aligned} \tilde{N}_1 - \tilde{N}_2 &= (\alpha(N_1 * \mathcal{S}_c)N_1 * \mathcal{S}_c - \alpha(N_2 * \mathcal{S}_c)N_2 * \mathcal{S}_c) \\ &\quad + \left\{\gamma(N_1 * \mathcal{S}_c)(1 - k(N_1 * \mathcal{S}_c))\left(2v_1 - \frac{v_1^2}{N_1}\right) * \mathcal{S}_c\right. \\ &\quad \left. -\gamma(N_2 * \mathcal{S}_c)(1 - k(N_1 * \mathcal{S}_c))\left(2v_2 - \frac{v_2^2}{N_2}\right) * \mathcal{S}_c\right\} \\ &= I + II + III, \end{aligned}$$

where

$$I = \alpha(N_1 * \mathcal{S}_c)N_1 * \mathcal{S}_c - \alpha(N_2 * \mathcal{S}_c)N_2 * \mathcal{S}_c,$$

$$\mathit{II} = [\gamma(N_1 * \mathcal{S}_c)(1 - k(N_1 * \mathcal{S}_c)) - r(N_2 * \mathcal{S}_c)(1 - k(N_2 * \mathcal{S}_c))]\left(2v_1 - \frac{v_1^2}{N_1}\right) * \mathcal{S}_c,$$

$$\mathit{III} = \gamma(N_2 * \mathcal{S}_c)(1 - k(N_2 * \mathcal{S}_c))\left[2(v_1 - v_2) - \left(\frac{v_1^2}{N_1} - \frac{v_2^2}{N_2}\right)\right] * \mathcal{S}_c.$$

By the Mean Value Theorem,

$$|I| \le (\alpha'(N_1^*)N_1^* + 1)|N_2 - N_1| * \mathcal{S}_c.$$

Similarly,

$$|\mathit{II}| \le [(\max_{N_1^* \le N \le N^*} |\gamma'(N)|)(1 - \kappa) + \gamma(N_1^*)\eta](3N^*)|N_1 - N_2| * \mathcal{S}_c,$$

$$|\mathit{III}| \le \gamma(N_1^*)(1 - \kappa)\left[\left(3 + \frac{N^*}{N_1^*}\right)|v_1 - v_2| * \mathcal{S}_c + \frac{N^*}{N_1^*}|N_1 - N_2| * \mathcal{S}_c\right].$$

Thus, for all $x \in \mathbb{R}$,

$$|\tilde{N}_1(x) - \tilde{N}_2(x)| \le \rho_1 \max\{|v_1 - v_2| * \mathcal{S}_c(x), |N_1 - N_2| * \mathcal{S}_c(x)\},$$

where

$$\rho_1 = \rho_1(\kappa, \eta) = (\alpha'(N_1^*)N_1^* + 1) + (3\gamma(N_1^*)N^*)\eta + (1 - \kappa)\left[3N^*\left(\max_{N_1^* \le N \le N^*} |\gamma'(N)|\right) + \gamma(N_1^*)\left(3 + 2\frac{N^*}{N^*}\right)\right].$$

Note that

$$\lim_{\kappa \to 1, \eta \to 0} \rho_1(\kappa, \eta) = \alpha'(N_1^*)N_1^* + 1 < 1.$$

We can choose $k_1 \ge k_0$ and $0 \le \eta_1 \le \eta_0$ such that $\rho_1(\kappa, \eta) < 1$ for $k_1 \le \kappa$ and $0 \le \eta \le \eta_1$. The choice of k_1 and η_1 is independent of c. By the definition of the metric $\boldsymbol{d}$, we have to estimate $\|\tilde{R}_1 - \tilde{R}_2\|_\infty$, $\|\tilde{V}_1 - \tilde{V}_2\|_\infty$ and $\|\tilde{U}_1 - \tilde{U}_2\|_\infty$. By

$$\tilde{R}_1(x) - \tilde{R}_2(x) = (Q_{1c}[N_1, v_1](x) - Q_{2c}[N_2, v_2](x))e^{(\sigma+\epsilon)x},$$

we get

$$\begin{aligned}
|\tilde{R}_1(x) - \tilde{R}_2(x)| &\le \rho_1 \max\{|v_1 - v_2| * \mathcal{S}_c(x), |N_1 - N_2| * \mathcal{S}_c(x)\}e^{(\sigma+\epsilon)x} \\
&\le \rho_1 e^{-(\sigma+\epsilon)c}K(\sigma + \epsilon) \max\{|V_1 - V_2| * F_1(x), |R_1 - R_2| * F_1(x)\} \\
&\le \rho_1 \max\{\|V_1 - V_2\|_\infty, \|R_1 - R_2\|_\infty\} \\
&\le \rho_1 \boldsymbol{d}([R_1, V_1], [R_2, V_2]),
\end{aligned}$$

where

$$dF_1(x) = (e^{-(\sigma+\epsilon)c}K(\sigma + \epsilon))^{-1}e^{(\sigma+\epsilon)x}d\mathcal{S}_c(x).$$

and we have used

$$e^{-(\sigma+\epsilon)c}K(\sigma + \epsilon) < \frac{1}{\gamma(N_1^*)} < 1.$$

Similarly, from Lemma 5.3, we get

$$\begin{aligned}|\tilde{V}_1(x) - \tilde{V}_2(x)| &\le \gamma(N_1^*) \max\{|v_1 - v_2| * \mathcal{S}_c(x), |u_1 - u_2| * \mathcal{S}_c(x)\} e^{(\sigma+\epsilon)x} \\ &= \gamma(N_1^*) e^{-(\sigma+\epsilon)c} K(\sigma+\epsilon) \max\{|V_1 - V_2| * F_1, |U_1 - U_2| * F_1\} \\ &\le \gamma(N_1^*) e^{-(\sigma+\epsilon)c} K(\sigma+\epsilon)\, \boldsymbol{d}([R_1, V_1], [R_2, V_2]),\end{aligned}$$

and

$$\begin{aligned}|\tilde{U}_1(x) - \tilde{U}_2(x)| &\le \gamma(N_1^*) \max\{|v_1 - v_2|, |u_1 - u_2|, |N_1 - N_2|\} * \mathcal{S}_c(x)\, e^{(\sigma+\epsilon)x} \\ &= \gamma(N_1^*) e^{-(\sigma+\epsilon)c} K(\sigma+\epsilon) \max\{|V_1 - V_2|, |U_1 - U_2|, |R_1 - R_2|\} * F_1 \\ &\le \gamma(N_1^*) e^{-(\sigma+\epsilon)c} K(\sigma+\epsilon)\, \boldsymbol{d}([R_1, V_1], [R_2, V_2]).\end{aligned}$$

Therefore

$$\boldsymbol{d}([\tilde{R}_1, \tilde{V}_1], [\tilde{R}_1, \tilde{V}_2]) \le \rho \boldsymbol{d}([R_1, V_1], [R_2, V_2]),$$

where

$$\rho = \max\{\rho_1, \gamma(N_1^*) e^{-(\sigma+\epsilon)c} K(\sigma+\epsilon)\} < 1.$$

Hence, whenever $k_1 \le \kappa$ and $0 \le \eta \le \eta_1$, $\Gamma : \mathcal{B} \to \mathcal{B}$ is a contraction mapping. Thus Γ has a unique fixed point $[R, V]$. The corresponding $[N(x), v(x)]$ satisfies

$$[N(x), v(x)] = Q_c[N, v](x)$$

with $N_1^* \le N(x) \le N^*$, $0 \le v(x) \le N(x)$ for all $x \in \mathbb{R}$, and $N(\infty) = N_1^*$, $v(\infty) = 0$. Since $R(x)$ and $V(x)$ are bounded, $N(x)$ and $v(x)$ are not constant functions. We *claim* that $N(-\infty)$ and $v(-\infty)$ exist with $N(-\infty) = v(-\infty) = N^*$. If the claim is proved, $[N(x), v(x)]$ is the desired traveling wave solution.

To prove the claim, we note that $\gamma(N * \mathcal{S}_c) \ge 1$ for all $x \in \mathbb{R}$. Thus $v(x) \ge v(x) * \mathcal{S}_c(x)$ for $x \in \mathbb{R}$. Since $v(x)$ is bounded, Theorem 4.1 of Essén [3, p. 126] implies that $v(-\infty)$ exists and $v(-\infty) \ne 0$. However,

$$\gamma(N * \mathcal{S}_c(x)) = \frac{v(x)}{v * \mathcal{S}_c(x)} \quad \text{for } x \in \mathbb{R}.$$

We have the limit $\lim_{x\to-\infty} \gamma(N * \mathcal{S}_c(x)) = \frac{v(-\infty)}{v(-\infty)} = 1$. Since $\gamma(N)$ is strictly decreasing in N, we have $\lim_{x\to-\infty} N * \mathcal{S}_c(x) = N^*$. By the General Tauberian Theorem of Wiener [18], we deduce that

$$\lim_{x\to\infty} N * h(x) = N^* \int_{-\infty}^{\infty} h(x)\, dx$$

for all $h(x) \in L^1(\mathbb{R})$. Thus Theorem 9 in Chapter V of Widder [17] implies that $N(-\infty)$ exists and $N(-\infty) = N^*$. From the N equation, we get (by letting $x \to -\infty$)

$$N^* = \gamma(N^*)N^* - (1 - \alpha(N^*)) \frac{(N^* - v(-\infty))^2}{N^*}.$$

Therefore $v(-\infty) = N^*$. This proves the claim.

The uniqueness of this traveling wave for $c > c_1^*$ follows from Lemma 5.1 (c) and the uniqueness of the fixed point of Γ.

We now prove the existence of a traveling wave for $c = c_1^*$. Define

$$c_n = c_1^* + \frac{1}{n}, \qquad n = 1, 2, 3, \ldots.$$

For each c_n, there exists a unique traveling wave $[N_n(x), v_n(x)]$ of the system Q with speed c_n, which satisfies the following conditions:

$$\begin{aligned} &N_n(-\infty) = N^*, \quad N_n(\infty) = N_1^*, \quad N_1^* \le N_n(x) \le N^*, \\ &v_n(-\infty) = N^*, \quad v_n(\infty) = 0, \quad 0 \le v_n(x) \le N_n(x), \end{aligned}$$

for all $x \in \mathbb{R}$. And we normalize (by translation) each wave such that

$$v_n(0) = \frac{1}{2}N^*, \qquad \frac{1}{2}N^* \le v_n(x) \le N^*$$

for all $x \le 0$. The sequence of functions $[N_n(x+c_n), v_n(x+c_n)]$ is uniformly bounded. Since $\mathcal{S}$ has a bounded probability density, the sequence of functions $[N_n(x + c_n) * \mathcal{S}(x), v_n(x + c_n) * \mathcal{S}(x)]$ is equicontinuous on every compact subset of $\mathbb{R}$. But

$$N_n(x + c_n) * \mathcal{S}(x) = \int_{-\infty}^{\infty} N_n(x - y + c_n) d\mathcal{S}(y) = N_n * \mathcal{S}_{c_n}(x).$$

Thus $[N_n * \mathcal{S}_{c_n}, v_n * \mathcal{S}_{c_n}]$ are equicontinuous on every compact subset of $\mathbb{R}$. By $N_1^* \le N_n(x)$, $v_n(x)^2/N_n(x)$ are uniformly bounded in $\mathbb{R}$. Therefore $(v_n^2/N_n) * \mathcal{S}_{c_n}(x)$ are equicontinuous on compact subsets of $\mathbb{R}$ also. Since

$$\begin{aligned} N_n &= \alpha(N_n * \mathcal{S}_{c_n}) N_n * \mathcal{S}_{c_n} + (\gamma(N_n * \mathcal{S}_{c_n}) - \alpha(N_n * \mathcal{S}_{c_n})) \left\{ 2v_n - \frac{v_n^2}{N_n} \right\} * \mathcal{S}_{c_n}, \\ v_n &= \gamma(N_n * \mathcal{S}_{c_n}) v_n * \mathcal{S}_{c_n}, \end{aligned}$$

it follows that $[N_n, v_n]$ are equicontinuous on compact subsets of $\mathbb{R}$. Thus we can extract a subsequence of $[N_n, v_n]$ which converges to a limit $[N, v]$ uniformly on every compact subset of $\mathbb{R}$. For convenience, we may assume that $\lim_{n\to\infty} N_n(x) = N(x)$ and $\lim_{x\to\infty} v_n(x) = v(x)$. Since this convergence is uniform on compact subsets of $\mathbb{R}$ and c_n is a bounded sequence, the limits

$$\lim_{n\to\infty} N_n(x + c_n) = N(x + c_1^*), \qquad \lim_{n\to\infty} v_n(x + c_n) = v(x + c_1^*)$$

hold for all $x \in \mathbb{R}$. Thus, by the Lebesgue Dominated Convergence Theorem,

$$\lim_{n\to\infty} N_n * \mathcal{S}_{c_n}(x) = N * \mathcal{S}_{c_1^*}(x), \qquad \lim_{n\to\infty} v_n * \mathcal{S}_{c_n}(x) = v * \mathcal{S}_{c_1^*}(x)$$

for all $x \in \mathbb{R}$. Then $[N(x), v(x)]$ satisfies

$$[N(x), v(x)] = Q_{c_1^*}[N, v](x), \quad N_1^* \le N(x) \le N^*, \quad 0 \le v(x) \le N(x)$$

for all $x \in \mathbb{R}$. Moreover, $v(0) = \frac{1}{2}N^*$ and $\frac{1}{2}N^* \le v(x) \le N^*$ for all $x \le 0$. It remains to show that $[N, v]$ is the desired traveling wave, i.e., $[N(\pm\infty), v(\pm\infty)]$ exist and they satisfy the boundary conditions (7).

By using $\gamma(N * \mathcal{S}_{c_1^*}(x)) \ge 1$ for all $x \in \mathbb{R}$ and the v equation, we get

$$v(x) \ge v * \mathcal{S}_{c_1^*}(x) \quad \text{for all } x \in \mathbb{R}.$$

Since $v(x)$ is bounded, Theorem 4.1 of Essén [3, p. 126] implies that $v(\pm\infty)$ exist. If $v(\infty) = v(-\infty)$, then $v(x)$ is a constant which must be $\frac{1}{2}N^*$. From the v equation, we get $\gamma(N * \mathcal{S}_{c_1^*}(x)) = 1$ for all $x \in \mathbb{R}$. Thus $N(x) = N^*$ for all $x \in \mathbb{R}$. But $[N^*, \frac{1}{2}N^*]$ can not satisfy the N equation. This contradiction establishes that $v(\infty) \neq v(-\infty)$. Now, from $\frac{1}{2}N^* \leq v(x) \leq N^*$ for $x \leq 0$, we get $v(-\infty) \geq \frac{1}{2}N^*$. Hence

$$\lim_{x\to-\infty} \gamma(N * \mathcal{S}_{c_1^*}(x)) = \lim_{x\to\infty} \frac{v(x)}{v * \mathcal{S}_{c_1^*}(x)} = \frac{v(-\infty)}{v(-\infty)} = 1.$$

We find as before that $N(-\infty) = N^*$. Because $[N(-\infty), v(-\infty)]$ is a constant fixed point of Q, the N equation implies that $v(-\infty) = N^*$. Thus $v(\infty) \neq N^*$. If $v(\infty) > 0$, then the same argument implies that $N(\infty) = N^*$, and hence $v(\infty) = N^*$. This is a contradiction. Therefore we have proved that $v(\infty) = 0$. Finally, we need to show that $N(\infty)$ exists and is N_1^*. For this purpose, we define $a = \limsup_{x\to\infty} N(x)$, $N_1^* \leq a \leq N^*$. We recall that the N equation has the form

$$N = \alpha(N * \mathcal{S}_{c_1^*})N * \mathcal{S}_{c_1^*} + E[N, v],$$

where

$$|E[N,v]| = \left|(\gamma(N * \mathcal{S}_{c_1^*}) - \alpha(N * \mathcal{S}_{c_1^*}))\left(2v - \frac{v^2}{N}\right) * \mathcal{S}_{c_1^*}\right| \leq \gamma(N_1^*)\left(2 + \frac{N^*}{N_1^*}\right) v * \mathcal{S}_{c_1^*}.$$

Thus $\lim_{x\to\infty} E[N,v](x) = 0$. By taking the limit sup in the N equation as $x \to \infty$, we get $a \leq \alpha(a)a$. Hence $\alpha(a) \geq 1$. However, $N_1^* \leq a$, $\alpha(a) \leq \alpha(N_1^*) = 1$. Therefore $\alpha(a) = 1$, i.e., $a = N_1^*$. But $N_1^* \leq N(x) \leq N^*$, we have $\liminf_{x\to\infty} N(x) \geq N_1^* = \limsup_{x\to\infty} N(x)$. Thus $N(\infty)$ exists and $N(\infty) = N_1^*$. This proves Theorem 2.2. □

We remark that the above proof used the following facts from simple analysis. Let $f(x)$ be a bounded continuous function and g be a continuous nondecreasing function. Suppose that F is a probability distribution over $\mathbb{R}$. Then

$$\limsup_{x\to\infty} f * F(x) \leq \limsup_{x\to\infty} f(x), \qquad \limsup_{x\to\infty} g(f(x)) = g(\limsup_{x\to\infty} f(x)).$$

These are easily proved by using the definition of limit superior.

6. The proof of Theorem 2.3. We will modify the proof of Theorem 2.2 to show the asymptotic behavior in Theorem 2.3. As in §5, we will use $[N, v]$ instead of $[N, u]$. The mapping Q_c is defined in §5.

Let $v^{(0)}(x) = N^{(0)}(x) - u^{(0)}(x) = e^{-\sigma x}\{B^{(0)} + e^{-\epsilon x}V^{(0)}(x)\}$ where $B^{(0)} = A^{(0)} - C^{(0)}$ and $V^{(0)}(x) = R^{(0)}(x) - U^{(0)}(x)$. Define the sequence $[\bar{N}^{(n)}(x), \bar{v}^{(n)}(x)]$ as follows:

$$[\bar{N}^{(n+1)}(x), \bar{v}^{(n+1)}(x)] = Q_c[\bar{N}^{(n)}, \bar{v}^{(n)}](x), \qquad n = 0, 1, 2, \ldots,$$
$$[\bar{N}^{(0)}(x), \bar{v}^{(0)}(x)] = [N^{(0)}(x), u^{(0)}(x)].$$

Then we prove by induction that

$$[\bar{N}^{(n)}(x), \bar{v}^{(n)}(x)] = [N^{(n)}(x + nc), v^{(n)}(x + nc)] \quad \text{for } x \in \mathbb{R},\ n = 0, 1, 2, \ldots.$$

Moreover, by Lemma 5.2, we have

$$\bar{N}^{(n)}(x) = L^{(n)} + e^{-\sigma x}\{A^{(n)} + e^{-\epsilon x}R^{(n)}(x)\},$$
$$\bar{v}^{(n)}(x) = e^{-\sigma x}\{B^{(n)} + e^{-\epsilon x}V^{(n)}(x)\}.$$

$L^{(n)}$, $A^{(n)}$, and $B^{(n)}$ satisfy the recursive relations in Lemma 5.2. $R^{(n)}(x)$ and $V^{(n)}(x)$ are bounded continuous functions.

We choose η_1 such that, when $\max_{0\le N\le N^*}\{|1-k(N)|, |k'(N)|\} \le \eta_1$, we have $\alpha'(N^*)N^*+\alpha(N^*) > 0$ (due to (A4)'). Then, since $0 < L^{(0)} < 1$, we get $\lim_{n\to\infty} L^{(n)} = N_1^*$. The rate of convergence is exponential in n. If σ and c satisfy the relation $e^{-\sigma c}K(\sigma)\gamma(N_1^*) = 1$, then the infinite product

$$L^{(\infty)} = \prod_{n=0}^{\infty}[e^{-\sigma c}K(\sigma)\gamma(L^{(n)})]$$

converges and $L^{(\infty)} > 0$. The sequences $B^{(n)}$ and $A^{(n)}$ converge to the limits $B^{(\infty)}$ and $A^{(\infty)}$, respectively, where

$$B^{(\infty)} = L^{(\infty)}B^{(0)}, \qquad A^{(\infty)} = \frac{2(\gamma(N_1^*)-1)L^{(\infty)}}{\gamma(N_1^*)-(\alpha'(N_1^*)N_1^*+1)}B^{(0)}.$$

The rate of convergence is also exponential.

We *claim* that

$$N_1^* \le \liminf_{n\to\infty} \bar{N}^{(n)}(x) \le \limsup_{n\to\infty} \bar{N}^{(n)}(x) \le N^* \tag{12}$$

uniformly in $x \in \mathbb{R}$. That $\limsup_{n\to\infty} \bar{N}^{(n)}(x) \le N^*$ follows from Step 2 of the proof of Theorem 2.1. Due to the choice of η_1, there exists $\delta > 0$ and n_0 such that $\sup_{x\in\mathbb{R}} \bar{N}^{(n)}(x) \le N^* + \delta$ for all $n \ge n_0$, and $N\alpha(N)$ is strictly increasing in $0 \le N \le N^* + \delta$. We define the sequence $\hat{N}^{(n)}(x)$ as follows:

$$\hat{N}^{(n+1)} = \alpha(\hat{N}^{(n)} * \mathcal{S}_c)\hat{N}^{(n)} * \mathcal{S}_c, \qquad n = n_0, n_0+1, n_0+2, \dots,$$
$$\hat{N}^{(n_0)}(x) = \bar{N}^{(n_0)}(x).$$

By (A2)', $Q_{1c}[N,u] \ge \alpha(N * \mathcal{S}_c)N * \mathcal{S}_c$. An induction argument gives that $\bar{N}^{(n)}(x) \ge \hat{N}^{(n)}(x)$ for $n \ge n_0$ and $x \in \mathbb{R}$. If $N^{(0)}(\pm\infty) > 0$, Step 1 of the proof of Theorem 2.1 implies that $\bar{N}^{(n)}(\pm\infty) > 0$ for all n. But $\alpha'(N_1^*)N_1^* + 1 > 0$ (due to the choice of η_1), and Theorem 3 of Weinberger [15] implies that

$$\lim_{n\to\infty} \hat{N}^{(n)}(x) = N_1^* \quad \text{uniformly in } \mathbb{R}.$$

Hence

$$\liminf_{n\to\infty} \bar{N}^{(n)}(x) \ge N_1^* \quad \text{uniformly in } \mathbb{R}.$$

The claim is proved.

We prove Theorem 2.3 (b), (c) first. By the assumption, $0 < \sigma < \lambda^*$. Let $c = \Psi(\sigma)$. Then $e^{-\sigma c}K(\sigma)\gamma(N_1^*) = 1$. Thus $L^{(\infty)}$, $A^{(\infty)}$ and $B^{(\infty)}$ are finite numbers. We can choose $0 < \eta_2 \le \eta_1$ and $\delta_1 > 0$ such that, when $\max_{0\le N\le N^*}\{|1-k(N)|, |k'(N)|\} \le \eta_2$, we have

$$\gamma(N_1^*-\delta)e^{-(\sigma+\epsilon)c}K(\sigma+\epsilon) < 1, \qquad \alpha'(N_1^*-\delta)(N_1^*-\delta)+\alpha(N_1^*-\delta) < 1$$

for all $0 < \delta \le \delta_1$. The space $\mathcal{B}$ in the proof of Theorem 2.2 is modified as follows:

$$\mathcal{B} = \{[L, A, B, R(x), V(x)] \mid |L-N_1^*| \le \delta,\ |A-A^{(\infty)}| \le \delta,\ |B-B^{(\infty)}| \le \delta.$$
$R(x)$ and $V(x)$ are bounded continuous functions in $\mathbb{R}$. Let
$N(x) = L + e^{-\sigma x}[A + e^{-\epsilon x}R(x)]$ and $v(x) = e^{-\sigma x}[B + e^{-\epsilon x}V(x)]$.
$0 \le v(x) \le N(x)$ and $N_1^* - \delta \le N(x) \le N^* + \delta$ for all $x \in \mathbb{R}\}$.

For $[L_i, A_i, B_i, R_i, V_i] \in \mathcal{B}$, $i = 1, 2$, the metric $\boldsymbol{d}$ is defined as

$$\begin{aligned}&\boldsymbol{d}([L_1, A_1, B_1, R_1, V_1], [L_2, A_2, B_2, R_2, V_2])\\&\quad = \max\{|L_1 - L_2|, |A_1 - A_2|, |B_1 - B_2|, \|R_1 - R_2\|_\infty, \|V_1 - V_2\|_\infty, \|U_1 - U_2\|_\infty\},\end{aligned}$$

where $U_i = R_i - V_i$, $i = 1, 2$. $(\mathcal{B}, \boldsymbol{d})$ is again a complete metric space. The mapping Q_c now corresponds to a mapping Γ defined by

$$\Gamma[L, A, B, R, V] = [\tilde{L}, \tilde{A}, \tilde{B}, \tilde{R}, \tilde{V}],$$

where the image is given as in Lemma 5.2. Since $\alpha'(N_1^*)N_1^* + 1 > 0$, $\delta > 0$ can be chosen such that $\Gamma : \mathcal{B} \to \mathcal{B}$ is assured. A Lipschitz estimate on $R(x)$ and $V(x)$ similar to Lemma 5.3 establishes the contraction property of Γ. Indeed, $0 < \eta \le \eta_2$ can be chosen such that, whenever $\max_{0 \le N \le N^*}\{|1 - k(N)|, |k'(N)|\} \le \eta$, the following estimate holds:

$$\begin{aligned}&\boldsymbol{d}(\Gamma[L_1, A_1, B_1, R_1, V_1], \Gamma[L_2, A_2, B_2, R_2, V_2])\\&\quad \le \rho\, \boldsymbol{d}([L_1, A_1, B_1, R_1, V_1], [L_2, A_2, B_2, R_2, V_2]),\end{aligned}$$

where

$$\rho = \max\{\gamma(N_1^* - \delta)e^{-(\sigma+\epsilon)c}K(\sigma + \epsilon),\ \ \alpha'(N_1^* - \delta)(N_1^* - \delta) + \alpha(N_1^* - \delta)\}.$$

Therefore Γ has a unique fixed point in $\mathcal{B}$. We shall identify this fixed point in cases (b) and (c), respectively.

By the claim (12), we can find n_1 such that $N_1^* - \delta \le \bar{N}^{(n)}(x) \le N^* + \delta$ for $x \in \mathbb{R}$, $|L^{(n)} - N_1^*| \le \delta$, $|A^{(n)} - A^{(\infty)}| \le \delta$, and $|B^{(n)} - B^{(\infty)}| \le \delta$ for all $n \ge n_1$. Thus $[L^{(n)}, A^{(n)}, B^{(n)}, R^{(n)}(x), V^{(n)}(x)] \in \mathcal{B}$ for $n \ge n_1$.

(b) When $A^{(0)} = C^{(0)}$, we have $B^{(n)} = B^{(0)} = 0$ for all n. Hence $A^{(\infty)} = B^{(\infty)} = 0$. Then $[N_1^*, 0, 0, 0, 0] \in \mathcal{B}$ and is the fixed point of Γ. That is,

$$\begin{aligned}&\lim_{n\to\infty} \sup_{x \ge l} |\bar{N}^{(n)}(x) - N_1^*|\\&= \lim_{n\to\infty} \sup_{x \ge l} |(L^{(n)} - N_1^*) + e^{-\sigma x}A^{(n)} + R^{(n)}(x)e^{-(\sigma+\epsilon)x}| = 0,\end{aligned}$$

and

$$\lim_{n\to\infty} \sup_{x \in \mathbb{R}} \bar{v}^{(n)}(x)e^{(\sigma+\epsilon)x} = \lim_{n\to\infty} \sup_{x \in \mathbb{R}} V^{(n)}(x) = 0.$$

By $\bar{N}^{(n)}(x) = N^{(n)}(x + nc)$ and $\bar{v}^{(n)}(x) = v^{(n)}(x + nc)$, (b) is proved.

(c) When $A^{(0)} > C^{(0)}$, then $A^{(\infty)} > 0$. There exists a unique traveling wave $[N(x), v(x)]$ with speed c such that $\lim_{x\to\infty}(N(x) - N_1^*)e^{\sigma x} = A^{(\infty)}$. Define $R(x)$ and $V(x)$ as

$$N(x) = N_1^* + e^{-\sigma x}[A^{(\infty)} + R(x)e^{-\epsilon x}], \qquad v(x) = e^{-\sigma x}[B^{(\infty)} + V(x)e^{-\epsilon x}].$$

Then $[N_1^*, A^{(\infty)}, B^{(\infty)}, R(x), V(x)] \in \mathcal{B}$ and is the fixed point of Γ. Therefore

$$\lim_{n\to\infty} \|R^{(n)} - R\|_\infty = \lim_{n\to\infty} \|V^{(n)} - V\|_\infty = 0.$$

Hence

$$\lim_{n\to\infty}\sup_{x\geq l}|\bar{N}^{(n)}(x)-N(x)| = \lim_{n\to\infty}\sup_{x\geq l}|(L^{(n)}-N_1^*)+e^{-\sigma x}(A^{(n)}-A^{(\infty)}) + e^{-(\sigma+\epsilon)x}(R^{(n)}(x)-R(x))| = 0,$$

and

$$\lim_{n\to\infty}\sup_{x\geq l}|\bar{v}^{(n)}(x)-v(x)|e^{\sigma x} = \lim_{n\to\infty}\sup_{x\geq l}|(B^{(n)}-B^{(\infty)})+e^{-\epsilon x}(V^{(n)}(x)-V(x))| = 0.$$

Let $c' > c = \Psi(\sigma)$ be a larger wave speed. Then

$$\max_{x\geq nc'}|N^{(n)}(x)-N_1^*| = \sup_{y\geq n(c'-c)}|\bar{N}^{(n)}(y)-N_1^*| \leq \sup_{y\geq 0}|\bar{N}^{(n)}(y)-N(y)| + \sup_{y\geq n(c'-c)}|N(y)-N_1^*|.$$

But $\lim_{n\to\infty}\sup_{y\geq n(c'-c)}|N(y)-N_1^*| = 0$; therefore,

$$\lim_{n\to\infty}\max_{x\geq nc'}|N^{(n)}(x)-N_1^*| = 0.$$

(c) is proved.

(a) When $\sigma \geq \lambda^*$, then, for any $0 < \sigma_1 < \lambda^*$, we can write

$$N^{(0)}(x) = L^{(0)} + e^{-\sigma_1 x}\hat{R}(x)e^{-\epsilon_1 x}, \qquad u^{(0)}(x) = L^{(0)} + e^{-\sigma_1 x}\hat{U}(x)e^{-\epsilon_1 x},$$

where $0 < \epsilon_1 < \sigma - \sigma_1$. $\hat{R}$ and $\hat{U}$ are bounded continuous functions. Let $c_1 = \Psi(\sigma_1)$. Then (b) implies that

$$\lim_{n\to\infty}\max_{x\geq nc_1}|N^{(n)}(x)-N_1^*| = \lim_{n\to\infty}\max_{x\geq nc_1}|u^{(n)}(x)-N_1^*| = 0.$$

Since σ_1 can be chosen arbitrarily in $0 < \sigma_1 < \lambda^*$, $c_1 > c_1^*$ is arbitrary also. This proves (a).

The proof of Theorem 2.3 is complete. □

Remark. When $L^{(0)} = N_1^*$ and

$$A^{(0)} = \frac{2(\gamma(N_1^*)-1)}{\gamma(N_1^*)-1+\alpha'(N_1^*)N_1^*}C^{(0)}$$

in Theorem 2.3, the rate of convergence in (b) and (c) is exponential. If $C^{(0)} = 0$, then

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|N^{(n)}(x+nc)-N_1^*|e^{(\sigma+\epsilon)x} = \lim_{n\to\infty}\sup_{x\in\mathbb{R}}|u^{(n)}(x+nc)-N_1^*|e^{(\sigma+\epsilon)x} = 0.$$

On the other hand, if $C^{(0)} > 0$, then

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|N^{(n)}(x+nc)-N(x)|e^{(\sigma+\epsilon)x} = \lim_{n\to\infty}\sup_{x\in\mathbb{R}}|u^{(n)}(x+nc)-u(x)|e^{(\sigma+\epsilon)x} = 0.$$

7. Concluding remarks. We have shown some special cases of the dynamics of Q in Theorems 2.1–2.3 when the initial generation has positive local population density $N^{(0)}(x)$ at $x = \pm\infty$. In this section, we shall make a few remarks related to the case when $N^{(0)}(\pm\infty) = 0$ (for example, when $N^{(0)}(x)$ has compact support) and to the heterozygote superior case.

(a) *Traveling waves connecting* $[N^*, 0]$ *and* $[0, 0]$. We assume (A1)–(A4). $[N^*, 0]$ and $[0, 0]$ are always constant fixed points of Q. Define

$$c^* = \inf_{\lambda>0} \frac{1}{\lambda} \log(\gamma(0)K(\lambda)).$$

It is shown in Weinberger [15], [16] that the traveling wave equation

$$N = \gamma(N * \mathcal{S}_c)N * \mathcal{S}_c, \quad N(-\infty) = N^*, \quad N(\infty) = 0 \tag{13}$$

of the mapping $\gamma(N * \mathcal{S})N * \mathcal{S}$ admits a solution iff $c \geq c^*$. When $c > c^*$, the traveling wave is unique up to translation. Then, as a direct consequence of Remarks 3.1 and 3.2, *the system of the traveling wave equations*

$$[N(x - c), u(x - c)] = Q[N, u](x),$$
$$[N(-\infty), u(-\infty)] = [N^*, 0], \qquad [N(\infty), u(\infty)] = [0, 0]$$

admits a solution $[N(x), u(x)]$ *with* $\lim_{x\to\infty} \frac{u(x)}{N(x)} < 1$ *iff* $c \geq c^*$, *and the profile is given by* $[N(x), 0]$ *where* $N(x)$ *satisfies* (13). *When* $c > c^*$, *the wave* $[N(x), u(x)]$ *with* $\lim_{x\to\infty} \frac{u(x)}{N(x)} < 1$ *is also unique up to translation.* We note that, when $u(x) = 0$, the system of the traveling wave equations for Q becomes (13).

We try to remove the condition $\lim_{x\to\infty} \frac{u(x)}{N(x)} < 1$. Indeed, *if* $d\mathcal{P} = \delta$ *and* $\mathcal{S}$ *has a probability density of type* PF_2 (see Lui [11]), *then* $[N(x), 0]$, *where* $N(x)$ *satisfies* (13), *is the unique (up to translation) traveling wave of* Q *when* $c > c^*$.

The proof is as follows. Let $[\bar{N}, \bar{u}] \in \Delta$ be a traveling wave solution of the system with wave speed $c > c^*$. Define the sequence of functions $N^{(n)}(x)$ by

$$N^{(n+1)}(x) = \gamma(N^{(n)} * \mathcal{S}(x))\, N^{(n)} * \mathcal{S}(x), \quad N^{(0)}(x) = \bar{N}(x), \quad n = 0, 1, 2, \ldots.$$

By (A2), an induction argument shows that

$$0 \leq \bar{N}(x - nc) \leq N^{(n)}(x), \qquad n = 1, 2, 3, \ldots.$$

Since $\mathcal{S}$ has a probability density of type PF_2, $N^{(n)}(x)$ tends asymptotically to $N(x - nc + \xi)$ as $n \to \infty$ where ξ is some real number (see Lui [11]). Thus $0 \leq \bar{N}(x) \leq N(x + \xi)$ for all $x \in \mathbb{R}$. Let σ be the smallest positive root of $\gamma(0)e^{-\lambda c}K(\lambda) = 1$. Then $N(x)e^{\sigma x}$ is bounded, and so is $\bar{N}(x)e^{\sigma x}$. Since $0 \leq \bar{u}(x) \leq \bar{N}(x)$, $\bar{u}(x)e^{\sigma x}$ is also bounded.

We denote $R(x) = \bar{N}(x)e^{\sigma x}$ and $W(x) = \bar{u}(x)e^{\sigma x}$. By multiplying both sides of the traveling wave equations by $e^{\sigma x}$, we get

$$\begin{aligned} R = {} & \frac{\gamma(\bar{N} * \mathcal{S}_c)}{\gamma(0)} R * F + \frac{\alpha(\bar{N} * \mathcal{S}_c) - \gamma(\bar{N} * \mathcal{S}_c)}{\gamma(0)} \left(\frac{W^2}{R} * F\right) \\ & + 2\frac{\beta(\bar{N} * \mathcal{S}_c) - \gamma(\bar{N} * \mathcal{S}_c)}{\gamma(0)} \left(\frac{W(R - W)}{R} * F\right), \\ W = {} & \frac{\gamma(\bar{N} * \mathcal{S}_c)}{\gamma(0)} W * F + \frac{\alpha(\bar{N} * \mathcal{S}_c) - \gamma(\bar{N} * \mathcal{S}_c)}{\gamma(0)} \left(\frac{W^2}{R} * F\right) \\ & + \frac{\beta(\bar{N} * \mathcal{S}_c) - \gamma(\bar{N} * \mathcal{S}_c)}{\gamma(0)} \left(\frac{W(R - W)}{R} * F\right), \end{aligned} \tag{14}$$

where $dF(x) = \gamma(0)e^{\sigma x}d\mathcal{S}_c(x)$. By (A2) and $\gamma(\bar{N} * \mathcal{S}_c) \le \gamma(0)$, we have

$$0 \le R(x) \le R * F(x), \qquad 0 \le W \le W * F(x) \quad \text{for all } x \in \mathbb{R}.$$

Then Theorem 4.1 of Essén [3] implies that $R(\pm\infty)$ and $W(\pm\infty)$ exist. Since $R(-\infty) = 0$ and $R \not\equiv 0$, $R(\infty) > 0$. For convenience, let $A_1 = R(\infty)$ and $A_2 = W(\infty)$. Then, by taking $x \to \infty$ in (14), we get

$$A_1 = A_1 + \frac{(\alpha(0) - \gamma(0))}{\gamma(0)}\frac{A_2^2}{A_1} + \frac{2(\beta(0) - \gamma(0))}{\gamma(0)}\frac{A_2(A_1 - A_2)}{A_1},$$
$$A_2 = A_2 + \frac{(\alpha(0) - \gamma(0))}{\gamma(0)}\frac{A_2^2}{A_1} + \frac{(\beta(0) - \gamma(0))}{\gamma(0)}\frac{A_2(A_1 - A_2)}{A_1}.$$

By (A2), $\alpha(0) < \gamma(0)$. Thus $A_2 = 0$. We have shown that $W(\pm\infty) = 0$. Since $W(x) \le W * F(x)$ for all $x \in \mathbb{R}$, Theorem 4.1 of Essén [3] implies that $W(x) \equiv 0$. Thus $\bar{u}(x) \equiv 0$. Therefore

$$\bar{N}(x) = \gamma(\bar{N} * \mathcal{S}_c(x))\,\bar{N} * \mathcal{S}_c(x), \quad \bar{N}(-\infty) = N^*, \quad \bar{N}(\infty) = 0.$$

But $c > c^*$, and there exists some t such that $\bar{N}(x) = N(x + t)$.

The asymptotic behavior and the stability of the traveling waves of the scalar mapping $\gamma(N * \mathcal{S})N * \mathcal{S}$ are studied extensively in Weinberger [15], [16] and Lui [11]–[13]. They show that when the initial generation $N^{(0)}(x)$ satisfies $N^{(0)}(\infty) = 0$ in a certain sense, the nth generation $N^{(n)}(x)$ tends asymptotically to a traveling wave. If the initial generation $[N^{(0)}(x), u^{(0)}(x)]$ of the system Q satisfies $N^{(0)}(\infty) = 0$, we think that a similar form of Theorem 2.3 should hold when N_1^* is replaced by 0, and the traveling wave $[N(x), u(x)]$ is the wave connecting $[N^*, 0]$ and $[0, 0]$. In other words, c^* should be characterized as the asymptotic speed for the spread of the advantageous gene when $N^{(0)}(x)$ has compact support.

(b) *The degenerate system Q where $\alpha(N) = \beta(N) = \gamma(N)$ for all $0 \le N \le 1$.* The system decouples into

$$\tilde{N} = \gamma(N * \mathcal{S})N * \mathcal{S}, \qquad \tilde{u} = \frac{1}{2}\gamma(N * \mathcal{S})\left\{\left(u + N\frac{u * \mathcal{P}}{N * \mathcal{P}}\right) * \mathcal{S}\right\}.$$

Suppose that (A1), (A3), and (A4) hold. The constant fixed points consist of $[0, 0]$ and $[N^*, u^*]$ for all $0 \le u^* \le N^*$. If $[N^{(0)}(x), u^{(0)}(x)] \in \Delta$ satisfies $N^{(0)}(\pm\infty) > 0$, then $N^{(n)}(x)$ tends to N^* uniformly in $\mathbb{R}$, and $u^{(n)}(x)$ tends to u^* uniformly on compact subsets of $\mathbb{R}$. u^* is given by

$$u^* = N^*\left\{(1 - j)\frac{u^{(0)}(-\infty)}{N^{(0)}(-\infty)} + j\frac{u^{(0)}(\infty)}{N^{(0)}(\infty)}\right\}, \qquad j = \lim_{n\to\infty} \underbrace{\mathcal{J} * \mathcal{J} * \cdots * \mathcal{J}}_{n \text{ times}},$$

where $\mathcal{J} = \frac{1}{2}(\mathcal{S} + \mathcal{P} * \mathcal{S})$. The system admits a traveling wave $[N(x), u(x)]$ with speed c, which satisfies the boundary conditions

$$[N(-\infty), u(-\infty)] = [N^*, u^*], \quad [N(\infty), u(\infty)] = [0, 0], \quad 0 \le u^* \le N^*$$

iff $c \ge c^*$. The wave is uniquely given by $u(x) = \frac{u^*}{N^*}N(x)$ where $N(x)$ satisfies (13). These waves are also the asymptotic states of some initial generations. These results will be published elsewhere, where the condition (A4) can be relaxed to include some cases of $-1 < \gamma'(N^*)N^* + 1 < 0$.

(c) *The heterozygote superior case.* When $\beta(N) > \max\{\alpha(N), \gamma(N)\}$ for all $0 \leq N < 1$, there exists the interior constant fixed point of Q, i.e., the so-called polymorphism, where two alleles coexist in equilibrium. The dynamics of Q restricted to constant states is already nontrivial. When the viability functions are density independent, the complete dynamics is known (for example, Karlin [6]). However, the mapping Q has density-dependent viability functions. One can show (see Lin [10]) that the dynamics of Q (restricted to constant states) tends asymptotically to the dynamics of the scalar mapping $NG(N)$, where

$$G(N) = \frac{\beta(N)^2 - \alpha(N)\gamma(N)}{2\beta(N) - \alpha(N) - \gamma(N)}.$$

The chaos occurs generically. In general, the relation between the system Q and the scalar mapping $G(N * \mathcal{S})N * \mathcal{S}$ is yet to be explored. Only some special cases where α, β, and γ satisfy certain relations are known. These results will be published elsewhere.

Acknowledgments. Most of the contents of this paper were part of the author's Ph.D. thesis [9]. The author would like to thank her advisor, Professor H. F. Weinberger, for his guidance. She also wishes to thank Professor W.-M. Ni for his concern and encouragement.

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Partial Differential Equations and Related Topics, Lecture Notes in Math., Vol. 446, Springer-Verlag, 1975, pp. 5–49.

[2] O. Diekmann and H. G. Kaper, *On the bounded solutions of a nonlinear convolution equation*, Nonlinear Anal. Theory, Methods, Appl., 2 (1978), pp. 721–737.

[3] M. Essén, *Studies on a convolution inequality*, Arkiv för Mathematik, 5 (1963), pp. 113–159.

[4] W. Feller, *An introduction to probability theory and its applications. Vol* II, John Wiley, New York, 1966.

[5] R. A. Fisher, *The advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 355–369.

[6] P. Kareiva and W. Morris et al., *Project "Experimental and theoretical studies of gene spread in weed populations,"* Dept. of Zoology, Univ. of Washington, Seattle, WA, funded by the USDA Weed Science Program, 1989.

[7] S. Karlin, *Theoretical aspects of multi-locus selection balance* I, Studies in Mathematical Biology, Part II: Population and Communities, S. A. Levin, ed., MAA Studies in Math., Vol. 16, 1978, pp. 503–587.

[8] A. Kolmogorov, I. Petrovsky, and N. Piscunoff, *Étude de l'équations de la diffusion avec croissance de la quantité de matière et son application a un probleme biologique*, Bull. Univ. Moscow, Ser. Internat., Sec. A, 1 (1937), pp. 1–25.

[9] H. T. Lin, *On the dynamics of a model in the propagation of genes*, Ph.D. thesis, University of Minnesota, Minneapolis, MN, 1991.

[10] ———, *A discrete one-locus two-allele density dependent selection model*, Bulletin of the Inst. of Math., Academia Sinica, 21 (1993), pp. 303–324.

[11] R. Lui, *A nonlinear integral operator arising from a model in population genetics*, I. *Monotone initial data*, SIAM J. Math. Anal., 13 (1982), pp. 913–937.

[12] ———, *A nonlinear integral operator arising from a model in population genetics*, II. *Initial data with compact support*, SIAM J. Math. Anal., 13 (1982), pp. 938–953.

[13] ———, *Existence and stability of traveling wave solutions of a non-linear integral operator*, J. Math. Biol., 16 (1983), pp. 199–220.

[14] E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Clarendon Press, Oxford, 1937.

[15] H. F. Weinberger, *Asymptotic behavior of a model in population genetics*, in Nonlinear Partial Differential Equations and Applications, J. Chadam, ed., Lecture Notes in Math. 648, Springer-Verlag, New York, 1978, pp. 47–96.

[16] H. F. WEINBERGER, *Long-time behavior of a class of biological models*, SIAM J. Math. Anal., 13 (1982), pp. 353–396.
[17] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1946.
[18] N. WIENER, *The Fourier Integral and Certain of its Applications*, Cambridge University Press, Cambridge, 1933.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 77–97, January 1995

004

BLOW-UP SOLUTIONS OF QUASI-LINEAR DEGENERATE PARABOLIC EQUATIONS WITH CONVECTION *

RYUICHI SUZUKI†

Abstract. The Cauchy problem of the quasi-linear degenerate parabolic equations with convection term is considered. Under some conditions, the existence of single-point blow-up solutions is shown; it is also shown that the blow-up point is bounded. In addition, the asymptotic behavior of interfaces of blow-up solutions is studied.

Key words. blow-up solution, single-point blow-up, quasi-linear degenerate parabolic equation, convection, asymptotic behaviors of interfaces, Cauchy problem

AMS subject classifications. 35B40, 35K15, 35K55, 35K65

0. Introduction.

In this paper we shall consider the Cauchy problem in $\mathbb{R}$:

$$\partial_t \beta(u) - u_{xx} + g(u)_x = f(u), \qquad (x,t) \in \mathbb{R} \times (0,T), \tag{0.1}$$

$$u(x,0) = \varphi(x), \qquad x \in \mathbb{R}, \tag{0.2}$$

where $\beta(v), g(v)$, and $f(v)$ with $v \geq 0$ and $\varphi(x)$ are nonnegative continuous functions.

Equation (0.1) describes the combustion process with convection in a stationary medium in which the thermal conductivity $\beta'(u)^{-1}$, the volume heat source $f(u)$, and convection $g(u)$ depend in a nonlinear way on the temperature $\beta(u) = \beta(u(x,t))$ of the medium.

Throughout this paper we assume

(A1) $\beta(v), f(v), g(v) \in C^\infty(\mathbb{R}_+) \cap C(\bar{\mathbb{R}}_+), \beta(v) > 0, \beta'(v) > 0, \beta''(v) \leq 0$, and $f(v), g(v), g'(v) > 0$ for $v > 0$; $\lim_{v\to\infty} \beta(v) = \infty$; $f \circ \beta^{-1}$ and $g \circ \beta^{-1}$ are locally Lipschitz continuous in $[\beta(0), \infty)$.

(A2) $\{g \circ \beta^{-1}\}'(u) \leq C\sqrt{\{\beta^{-1}\}'(u)}$ in the neighborhood of $u = 0$ for some positive constant C.

(A3) $\varphi(x) \geq 0, \not\equiv 0$, and $\in \mathbb{B}(\mathbb{R})$ (bounded continuous in $\mathbb{R}$).

With these conditions the above Cauchy problem has a unique local solution $u(x,t)$ (in time), which satisfies (0, 1) in $\mathbb{R} \times (0,T)$ in the following weak sense, where $T > 0$ is assumed to be sufficiently small (see, e.g., Oleinik et al. [8], [11], [15]).

DEFINITION 0.1. *Let G be an open interval in $\mathbb{R}$. By a weak solution of equation* (0.1) *in $G \times (0,T)$ we mean a function $u(x,t)$ such that*

(1) *$u(x,t) \geq 0$ in $\bar{G} \times [0,T)$ and $\in \mathbb{B}(\bar{G} \times [0,\tau])$ for each $0 < \tau < T$.*

(2) *For any bounded open interval $\Omega = (x_1, x_2) \subset G, 0 < \tau < T$ and $\varphi(x,t) \in C^2(\bar{\Omega} \times [0,T))$, which vanishes on $x = x_1, x_2$, the following identity holds:*

$$\begin{aligned} &\int_\Omega \beta(u(x,\tau))\varphi(x,\tau)\,dx - \int_\Omega \beta(u(x,0))\varphi(x,0)\,dx \\ &\quad = \int_0^\tau \int_\Omega \{\beta(u)\varphi_t + u\varphi_{xx} + g(u)\varphi_x + f(u)\varphi\}\,dx\,dt - \int_0^\tau u\varphi_x\,dt|_{x=x_1}^{x=x_2}. \end{aligned} \tag{0.3}$$

* Received by the editors August 24, 1992; accepted for publication (in revised form) September 14, 1993.

† Department of Mathematics, Tokyo Metropolitan College of Aeronautical Engineering, 8-52-1 Minamisenju, Arakawa-ku, Tokyo, 116 Japan.

Remark 0.2. If $2q \geq m+1$,

$$(u^{1/m})_t = u_{xx} - (u^{q/m})_x + u^{p/m} \qquad (m,p,q \geq 1) \tag{0.4}$$

satisfies (A1), (A2), and (A3).

If $u(x,t)$ does not exist globally in time, its existence time $T < \infty$ is defined by

$$T = \sup\{\tau > 0; u(x,t) \text{ is bounded in } \mathbb{R} \times [0,\tau]\}, \tag{0.5}$$

and we see that

$$\limsup_{t \uparrow T} \sup_{x \in \mathbb{R}} u(x,t) = \infty. \tag{0.6}$$

In this case we say that u is a blow-up solution and T is a blow-up time.

The main purpose of the present paper is the study of blow-up solutions. We are especially interested in the shape of the blow-up set which locates the "hot-spots" at the blow-up time. In addition, since our quasi-linear equation (0.1) has a property of the finite propagation of an interface, there are some interesting subjects such as asymptotic behavior of the interface near the blow-up time. These problems have been studied for (0.1) without convection term in [5], [6], [7], and [16], and by Mochizuki and Suzuki in [14].

First, we consider the finite propagation of interfaces of solutions of (0.1), (0.2). To deal with this we require the additional conditions:

(A4) $\varphi(x) > 0$ for $x \in (-a_1, a_1)$ and $= 0$ for $x \notin (-a_1, a_1)$,

(A5) $\lim_{v \to 0} \dfrac{g'}{\beta'}(v) = 0$,

(A6) $\beta(0) = f(0) = 0, \qquad \displaystyle\int_0^1 \frac{dv}{\beta(v)} < \infty.$

Remark 0.3. If $q > 1$ and $m > 1$, equation (0.4) satisfies (A5) and (A6).

Put

$$\Omega(t) = \{x \in \mathbb{R}; u(x,t) > 0\}, \qquad \Gamma(t) = \partial\Omega(t) \tag{0.7}$$

for each $t \in (0,T)$. Then the interface Γ is given by

$$\Gamma = \bigcup_{0 \leq t \leq T} \Gamma(t) \times \{t\}, \tag{0.8}$$

and under assumptions (A1)–(A6) we can show that $\Omega(t)$ is bounded and nondecreasing in $t \in [0,T)$ (see Theorem 1.7). Moreover, $\Omega(t)$ is represented by continuous functions $\xi_i(t) : [0,T) \to \mathbb{R}\,(i = 1,2)$ like $\Omega(t) = \{x | x \in (\xi_1(t), \xi_2(t))\}$. For the case without convection term in (0.1), these results have been shown by Knerr [11], Suzuki [16], and Mochizuki and Suzuki [14].

Next, we restrict ourselves to the blow-up solution of (0.1), (0.2) and shall study the shape of the blow-up set and the behavior of the interface of u near the blow-up time. The existence and nonexistence of a blow-up solution of (0.1), (0.2) is discussed in Friedman and Lacey [4], Imai and Mochizuki [9], and Imai, Mochizuki, and Suzuki

[10]. We assume the following condition, as given in [10], as a "necessary" condition to raise a blow-up:

$$\text{(A7)} \qquad \int_1^\infty \frac{\beta'(v)}{f(v)}\,dv < \infty.$$

Furthermore, we assume that $f(u)$ grows more rapidly than $g(u)$ and u (see (A10) in §3), and, for the initial data $\varphi(x)$, that

$$\text{(A8)} \qquad \varphi'' - \{g(\varphi)\}' + f(\varphi) \geq 0 \quad \text{in } \mathcal{D}',$$

$$\text{(A9)} \qquad \text{the lap-number of } \varphi(x) \text{ in } [a_1, a_2] \text{ is two.}$$

Here, we define the lap-number of $\varphi(x)$ in the following way.

DEFINITION 0.4 (see [13]). *Let $\bar{I} = [a, b]$ be a closed interval and $w = w(x)$ be a real-valued function on $[a, b]$. We say w is piecewise monotone if $\bar{I}$ can be divided into a finite number of nonoverlapping subintervals $J_1, J_2, \ldots, J_m (\cup_{i=1}^m J_i = \bar{I})$, on each of which w is monotone. Then there is the least value of the numbers m for which we can find a division $\{J_i\}$ as above. This value is called the lap-number of w on $[a, b]$ and is denoted by $l(w)$.*

Then, for the semilinear case $\beta(u) = u$, Friedman and Lacey [4] have shown the existence of single-point blow-up solutions of (0.1) for the Dirichlet problem. In this paper we extend this result to the Cauchy problem of a degenerate quasi-linear equation. Moreover, we can also get that the left side interfaces stay bounded as t tends to the blow-up time T. To state these results exactly, we need the definition of a blow-up set.

DEFINITION 0.5. *The blow-up set of u is defined as*

$$S = \{x \in \mathbb{R};\ \textit{there is a sequence } (x_i, t_i) \in \mathbb{R} \times (0, T) \\ \textit{such that } x_i \to x, t_i \to T,\ \textit{and } u(x_i, t_i) \to \infty \textit{ as } i \to \infty\},$$

and each $x \in S$ is called a blow-up point of u.

Our second result is that, if we add the assumption (A10) in §3, then we obtain $S = \{\eta_0\}$ for some $-\infty < \eta_0 \leq \infty$ and $-\infty < \lim_{t \uparrow T} \xi_1(t)$ (see Theorem 3.4).

The final question we consider is whether $\eta_0 < \infty$ or $\eta_0 = \infty$ holds. Our answer is that $\eta_0 < \infty$ (see Theorem 4.4) if we add another condition on f such that $f(u)$ grows more rapidly than $g(u)$ and u (see (A11) in §4), and if we choose a special initial data φ corresponding to (A11) (see (A12) in §4).

Remark 0.6. $p > \max\{m, 2q - m, m + q - 1\}$ and $2q \geq m + 1$; (0.4) satisfies (A10), (A11).

Remark 0.7. Condition (A2) is needed to show the uniqueness of weak solutions of (0.1). If uniqueness of weak solutions to (0.1) holds, the above results are valid without (A2).

The methods of proving these results are essentially the same as those in Friedman and Lacey [4] and Suzuki [16]. We use the smoothness and comparison principle and a property of the zero set of $u_x(x, t)$.

This paper is structured as follows. In §1 we summarize the above two principles and show the finite propagation of the interfaces of the solutions (see Theorem 1.7). In §2 we study the property of the zero set of $u_x(x, t)$, where the lap-number of the initial data of the solution $u(x, t)$ is 2. Using this property we prove the existence of single-point blow-up solutions in §3. Finally, adding some assumptions, we show that the blow-up point is bounded in §4.

1. A comparison principle and finite propagation of interfaces. In this section we begin with two propositions that will be fundamental tools in our study of the interfaces and the blow-up sets.

PROPOSITION 1.1 (smoothness principle). *Assume* (A1)–(A3). *Let G be an open interval and let u be a solution of* (0.1) *in $G \times (0,T)$ in the sense of Definition* 0.1. *If $u(\bar{x},\bar{t}) > 0$ for some $(\bar{x},\bar{t}) \in G \times (0,T)$, then u is a classical solution in a neighborhood W of $(\bar{x},\bar{t})$, and hence $u \in C^\infty(W)$.*

Proof. Note that $\beta(v), f(v), g(v) \in C^\infty(\mathbb{R}_+)$ and $\beta'(v) > 0$ for $v > 0$. Then the above proposition follows from the usual parabolic regularization method (see, e.g., Ladyzenskaja, Solonnikov, and Ural'ceva [12]). □

DEFINITION 1.2. *For each open interval $G \subset \mathbb{R}$, a supersolution (or subsolution) of* (0.1) *in $G \times (0,T)$ is defined by* (1) *and* (2) *of Definition* 0.1 *with equality in* (0.3) *replaced by $\geq$ (or $\leq$).*

PROPOSITION 1.3 (comparison principle). *Assume* (A1)–(A3). *Let u (or v) be a superposition (or subsolution) of* (0.1) in $G \times (0,T)$. *If $u \geq v$ on the parabolic boundary of $G \times (0,T)$, then we have $u \geq v$ in the whole $\bar{G} \times (0,T)$.*

Proof. See, e.g., Gilding [8]. □

Remark 1.4. Condition (A2) is required in the proof of Proposition 1.3. But this condition could have been replaced by a weaker condition if we added some regularity conditions on $\varphi(x)$ (see Diaz and Kersner [3]).

In the rest of this section, based on these principles, we shall show finite propagation of the interface in $t < T$. First, we prove several lemmas.

LEMMA 1.5 (positivity). *Assume* (A1)–(A3) *and* (A5). *Let u be a weak solution of* (0.1), (0.2). *If $u(\bar{x},\bar{t}) > 0$ for some $(\bar{x},\bar{t}) \in \mathbb{R} \times (0,T)$, then*

$$u(\bar{x},t) > 0 \quad \text{for } t \geq \bar{t}. \tag{1.1}$$

Proof. (cf. Suzuki [16] and Friedman and Lacey [4]). Without loss of generality we can assume $\bar{x} = 0$. Since u is continuous in $\mathbb{R} \times [0,T)$, there exist $a_0 > 0$ and $\delta > 0$ such that

$$u(x,t) \geq a_0 \quad \text{in } [-2\delta, 2\delta] \times [\bar{t}, \bar{t} + 2\delta).$$

Let $\rho(t)$ be the solution to

$$\rho'(t) = -\frac{\lambda\rho}{\beta'(\rho)} \quad \text{in } (\bar{t},\infty) \qquad \text{with } \rho(\bar{t}) = a, \tag{1.2}$$

where $\lambda = (\pi/2\delta)^2$ and $0 < a < a_0$. Integrating this, we have

$$\rho(t) = W^{-1}(W(a) - \lambda(t - \bar{t})), \quad \text{where } W(s) = \int_1^s \frac{\beta'(v)}{v}\, dv. \tag{1.3}$$

Note that $\beta'(v) > 0$ and $\beta''(v) \leq 0$ in $v > 0$. Then, as is easily seen, $W(s)$ is increasing in $s > 0$ and $W(s) \to -\infty$ as $s \downarrow 0$. Thus $\rho(t) > 0$ for each $t > \bar{t}$.

Now we put

$$v(x,t) = \rho(t) \sin \frac{\pi}{2\delta}(x - \delta). \tag{1.4}$$

Then, since $\beta'(\rho) \leq \beta'(v)$, we see

$$\beta(v)_t \leq v_{xx}, \qquad (x,t) \in (-\delta,\delta) \times (\bar{t},T). \tag{1.5}$$

Next we put

$$r(t) = \int_0^t \sup_{0\le\xi\le\rho(t)} \frac{g'}{\beta'}(\xi)\,dt. \tag{1.6}$$

Set

$$R_1 = \{0 < x < \delta, 0 < t < T\},$$
$$R_2 = \{r(t) - r(T) < x < 0, 0 < t < T\},$$

and

$$R_3 = \{r(t) - r(T) - \delta < x < r(t) - r(T), 0 < t < T\}.$$

In addition, we define a function $w(x,t)$ in the following way:

$$\begin{aligned} w(x,t) &= v(x,t) \quad \text{in } R_1,\\ w(x,t) &= \rho(t) \quad \text{in } R_2\\ w(x,t) &= v(x - r(t) + r(T), t) \quad \text{in } R_3\\ w(x,t) &= 0 \quad \text{otherwise.} \end{aligned}$$

Since

$$r(T) \le T \sup_{0\le\xi\le a} \frac{g'}{\beta'}(\xi), \tag{1.7}$$

then, from (A5), if $a > 0$ is small enough, we have

$$0 < r(T) < \delta.$$

Hence we obtain

$$\begin{aligned} w(x,\bar t) &= 0 \quad \text{in } x \le -2\delta, \quad x \ge 2\delta,\\ w(x,\bar t) &\le a \quad \text{in } -2\delta \le x \le 2\delta, \end{aligned}$$

and

$$w(x,\bar t) \le u(x,\bar t), \qquad x \in \mathbb{R}. \tag{1.8}$$

On the other hand, we compute

$$\begin{aligned} \beta(w)_t - w_{xx} + g(w)_x &= \beta(v)_t - v_{xx} + g'(v)v_x\\ &\le g'(v)v_x \le 0 \quad \text{in } R_1 \quad (\text{since } v_x \le 0 \text{ in } R_1),\\ \beta(w)_t - w_{xx} + g(w)_x &= \beta'(\rho(t))\rho'(t) \le 0 \quad \text{in } R_2 \quad (\text{since } \rho'(t) \le 0), \end{aligned}$$

and

$$\begin{aligned} \beta(w)_t - w_{xx} + g(w)_x &= \beta' v_t - \beta' r' v_x - v_{xx} + g'(v)v_x\\ &\le (g'(v) - r'\beta')v_x\\ &= \beta'(v)\left\{\frac{g'}{\beta'}(v) - \sup_{0\le\xi\le\rho(t)} \frac{g'}{\beta'}(\xi)\right\} v_x\\ &\le 0 \quad \text{in } R_3 \quad (\text{since } v \le \rho, v_x \ge 0 \text{ in } R_3 \text{ and (A5)}). \end{aligned}$$

By this computation we can see that $w(x,t)$ is a subsolution of (0.1) in $\mathbb{R}\times[0,T)$. Applying the comparison principle (Proposition 1.3) to w and u, we get

$$u(x,t)\geq w(x,t),\qquad (x,t)\in\mathbb{R}\times[\bar{t},T), \tag{1.9}$$

namely,

$$u(0,t)\geq\rho(t)>0,\qquad t\in[\bar{t},T). \tag{1.10}$$

□

This lemma and the comparison principle imply the existence of the interface; that is, if we put

$$\xi_1(t)=\inf\{x|u(x,t)>0\}, \tag{1.11}$$

$$\xi_2(t)=\sup\{x|u(x,t)>0\}, \tag{1.12}$$

and assume (A4), then we have

$$\{x|u(x,t)>0\}=(\xi_1(t),\xi_2(t))\quad\text{for each } t\in[0,T). \tag{1.13}$$

Furthermore, if we assume (A6), then we can obtain the finite propagation of the interface in $t<T$ using the following lemma.

LEMMA 1.6. *Assume* (A1)–(A6). *Let* $u(x,t)$ *be a weak solution of* (0.1), (0.2). *Suppose that there exist* $(a,t_1)\in\mathbb{R}\times[0,T)$ *and* $M>0$ *such that*

$$u(x,t_1)=0\quad \textit{for } x\geq a, \tag{1.14}$$

$$u(a,t)\leq M\quad \textit{for } t\in[t_1,T). \tag{1.15}$$

Then, there exist $l>0$ *and* $h>0$ *depending only on* M *such that*

$$u(x,t)=0\quad \textit{for } (x,t)\in[a+l,\infty)\times[t_1,t_1+h]\cap[t_1,T). \tag{1.16}$$

Furthermore, if $M>0$ *is small enough, we can take* $l>0$ *small enough.*

Proof. (cf. Lemma 2.2 in this paper, Lemma 2.3 of Mochizuki and Suzuki [14], and Knerr [11]). We construct a supersolution $w(x,t)$ of (0.1) in the form

$$w(x,t)=\psi^{-1}([\rho(t)-(x-a)]^+), \tag{1.17}$$

where $[g]^+=\max\{g,0\}$, $\psi(u)=\int_0^u dv/\beta(v)$, $\rho(t)=C(M)(t-t_1)+\psi(M)$, and $C(M)=1+\sup_{0\leq v\leq 2M}\{g'/\beta'+f/\beta\beta'\}$ (these functions are well defined since we have assumed (A5) and (A6)).

In fact, in the domain $\{\{x\geq a\}\times[t_1,t_1+k]\}\cap\{\rho(t)\geq x-a\}$, where $k=C(M)^{-1}\{\psi(2M)-\psi(M)\}$, we have $w=\psi^{-1}(\rho(t)-x+a)\leq\psi^{-1}(\rho(t_1+k))=2M$, and hence

$$\begin{aligned}&\frac{1}{\beta'(w)}\psi(w)_{xx}+|\psi(w)_x|^2-\frac{g'(w)}{\beta'(w)}\psi(w)_x+\frac{f(w)}{\beta(w)\beta'(w)}\\&\quad\leq 1+\frac{g'(w)}{\beta'(w)}+\frac{f(w)}{\beta(w)\beta'(w)}\leq C(M)=\partial_t\psi(w).\end{aligned} \tag{1.18}$$

Therefore, this w is extended by 0 to the whole $\{x \geq a\} \times [t_1, t_1 + k]$ as a supersolution of (0.1). To achieve this, we have only to note that $f(0) = 0$ and $\partial_x w(\rho(t) + a, t) = (\psi^{-1})'(0) = \beta(0) = 0$.

Moreover, we have

$$w(x, t_1) \geq 0 = u(x, t_1) \quad \text{on } x \geq a,$$
$$w(x, t) \geq \psi^{-1}(\rho(t)) \geq \psi^{-1}(\psi(M)) = M \geq u(x, t) \quad \text{on } x = a, \quad t \geq t_1.$$

Thus, Proposition 1.3 implies that

$$w(x, t) \geq u(x, t) \quad \text{in } \{x \geq a\} \times [t_1, t_1 + k'), \tag{1.19}$$

where $k' = \min\{k, T - t_1\}$.

By the property of $w(x, t)$, choosing $h = k'$ and $l = \rho(t_1 + h)$, we conclude the assertion of (1.16). Since $\psi(M) \to 0$ as $M \to 0$, we can choose k small if $M > 0$ is small enough. Hence we can choose $l > 0$ small enough also. □

THEOREM 1.7. *Assume* (A1)–(A6). *Let $u(x, t)$ be a weak solution to* (0.1), (0.2). *Then $\Omega(t)$ forms a bounded set in $\mathbb{R}$ and is nondecreasing in t:*

$$\Omega(t_1) \subset \Omega(t_2) \quad \text{if } t_1 < t_2,$$

and there exist continuous functions $\xi_i(t)\colon [0, T) \to \mathbb{R}$ $(i = 1, 2)$ such that

$$\Omega(t) = \{x | x \in (\xi_1(t), \xi_2(t))\}.$$

Proof. Proposition 1.3 and Lemmas 1.5 and 1.6 are reduced to Theorem 1.7 easily. □

2. The property of the zero set of $u_x(x, t)$. Throughout this section, we assume (A1)–(A6). Furthermore, we assume that the lap-number of the initial data $\varphi(x)$ is 2 (see (A9)). In this section we prove the next proposition.

PROPOSITION 2.1. *Let $u(x, t)$ and $\xi_i(t)$ be as in Theorem* 1.7. *If we assume* (A9), *then there exists a C^1-function $\eta(t)\colon (0, T) \to \mathbb{R}$ such that*

$$\{x \in (\xi_1(t), \xi_2(t)); u_x(x, t) = 0\} = \{\eta(t)\} \tag{2.1}$$

for each $t \in (0, T)$ and for some $\delta > 0$

$$-a_1 + \delta \leq \eta(t) \quad \text{for all } t \in (0, T). \tag{2.2}$$

First we give the following lemma.

LEMMA 2.2. *Let $\varphi_n(x)$ be a C^∞-function such that $\varphi_n(x) \geq 1/n, \varphi_n(\pm n) = 1/n$, and $\varphi_n(x)$ converges to $\varphi(x)$ as n goes to ∞ locally uniformly with respect to x. Furthermore, assume that the maximum point of $\varphi_n(x)$ is unique and $\sigma_a \varphi_n(x) \geq \varphi_n(x)$ in $x \leq a$ if $a \leq -a_1 + \delta$, and in $x \geq a$ if $a \geq a_1$ for some $\delta > 0$. Here we note that $(\sigma_a x + x)/2 = a$ and $\sigma_a u(x) = u(\sigma_a x)$. Let $u_n(x, t)$ be a classical solution of the initial boundary value problem*

$$\begin{aligned} &\partial_t \beta(u) - u_{xx} + g(u)_x = f(u), && (x, t) \in [-n, n] \times (0, T), \\ &u(x, 0) = \varphi_n(x), && x \in [-n, n], \\ &u(\pm n, t) = 1/n, && t > 0. \end{aligned} \tag{2.3}$$

Then $u_n(x,t) \geq 1/n$ *for* $(x,t) \in [-n,n] \times (0,T)$, *and* $u_n(x,t) \to u(x,t)$ *as* $n \to \infty$ *locally uniformly in* $\mathbb{R} \times [0,T)$.

Proof. See Gilding [8]. □

Remark 2.3. The existence of the above $\varphi_n(x)$ is guaranteed by assumptions (A4) and (A9).

LEMMA 2.4. *Let* $u_n(x,t)$ *be as in Lemma* 2.2. *Then for each* $T' \in (0,T)$ *there exists a* C^1*-function* $\eta_n(t) : (0,T') \to \mathbb{R}$ *with large* n *such that*

$$\{x \in (-n,n);\ u_{n,x}(x,t) = 0\} = \{\eta_n(t)\} \quad \text{for each } t \in (0,T'). \tag{2.4}$$

Furthermore,

$$-a_1 + \delta \leq \eta_n(t) \quad \text{for } t \in (0,T'), \tag{2.5}$$

where $\delta > 0$ *is as in Lemma* 2.2.

Before we show this lemma, we need some notation and definitions (cf. Chen and Matano [2] and Suzuki [16]).

Notation 2.5. Let $w(x)$ be a continuous real-valued function on K where K is a bounded closed interval in $\mathbb{R}$. We define the nodal number of w by

$$\nu_K(w) = \text{ the number of points } x \in K \text{ with } w(x) = 0.$$

DEFINITION 2.6. *We say that* $w \in C^1(K)$ *poses only simple zeros if* $w'(x) \neq 0$ *for any* $x \in K$ *such that* $w(x) = 0$. *The set of all such functions is denoted by* $\Sigma(K)$.

LEMMA 2.7 (see Angenent [1] and note Suzuki [16]). *Let* $p(x,t), q(x,t)$, *and* $r(x,t)$ *be locally bounded continuous functions on* $[a,b] \times (t_0,T)$ *with* $p_{xx}, p_{xt}, p_{tt}, p_x, p_t$, q_x, q_t *all locally bounded continuous. Furthermore, let* $p(x,t) > 0$ *and let* $w(x,t)$ *be a classical solution of*

$$w_t = p(x,t)w_{xx} + q(x,t)w_x + r(x,t)w, \qquad (x,t) \in [a,b] \times (t_0,T). \tag{2.6}$$

Assume that $w(a,t) \neq 0$ *and* $w(b,t) \neq 0$ *for any* $t \in (t_0,T)$. *Then*

(i) $\nu(w(\cdot,t))$ *is finite for any* $t \in (t_0,T)$ *and is monotone nonincreasing in* t;

(ii) *If* x_0 *is a multiple zero of* $w(\cdot,t_1)$, *then* $\nu(w(\cdot,t_2)) > \nu(w(\cdot,t_3))$ *for all* $t_0 < t_2 < t_1 < t_3 < T$;

(iii) *There exists a strictly decreasing sequence of points* $\{t_k\}$ *such that* $\{t_k\} \downarrow t_0$ *and* $w(x,t) \in \Sigma([a,b])$ *for any* $t \in (t_0,T) \setminus \{t_k\}$.

On the other hand, we have the following lemma about lap-number (see Matano [13]).

LEMMA 2.8 (Matano). *Let* $u(x,t)$ *be a solution of the following Dirichlet problem:*

$$\begin{aligned} &u_t = a(x,t)u_{xx} + b(x,t)u_x + f(t,u) && \text{in } [a,b] \times (0,T), \\ &u(x,0) = u_0(x) && \text{in } [a,b], \\ &u(a,t) = u(b,t) = 0 && \text{in } (0,T), \end{aligned}$$

where $u_0(x) \in C([a,b] \times [0,T)), a \in C^1([a,b] \times [0,T)), b \in C^\alpha([a,b] \times [0,T))$ *for some* $0 < \alpha < 1, f \in C^1([0,T) \times \mathbb{R})$, *and* $a(x,t) \geq \delta$ *in* $[a,b] \times [0,T)$ *for some* $\delta > 0$. *Then if we assume that* $u(x,t) \geq 0$ *in* $[a,b] \times [0,T)$ *and* $l(u(\cdot,0)) \neq 0$, *the lap-number* $l(u(\cdot,t))$ *is nonincreasing in* $t \in [0,T)$.

Proof of Lemma 2.4. Applying the maximum principle to $u_n(x,t)$, we obtain

$$u_n(x,t) \geq 1/n \quad \text{in } x \in [-n,n], \quad t > 0 \tag{2.7}$$

and

$$\pm u_x(\pm n, t) < 0 \quad \text{for } t > 0. \tag{2.8}$$

Note that the lap-number $l(u_n(\cdot, t)) = l(u_n(\cdot, t) - 1/n)$ for $t > 0$ is equal to two by Lemma 2.8. Hence, since the nodal number $\nu_{[-n,n]}(u(\cdot, t)) = 1$, it follows from Lemma 2.7 (ii) that there exists a C^1-function $\eta_n(t)$ such that

$$\{x \in (-n, n); u_{n,x}(x, t) = 0\} = \{\eta_n(t)\} \quad \text{for } t > 0. \tag{2.9}$$

Next we prove (2.5) (see Friedman and Lacey [4]). Choose $a \in [-n, -a_1 + \delta]$ and set

$$w = u(x, t) - v(x, t) \quad \text{in } [-n, a]$$

where $u = u_n$ and $v = \sigma_a u_n$. Then w satisfies

$$\beta'(u)w_t - w_{xx} + Cw = -\{g'(u) + g'(v)\}u_x + g'(v)w_x, \tag{2.10}$$

where

$$C = C(x, t) = -\frac{f(u) - f(v)}{u - v} + \frac{\beta'(u) - \beta'(v)}{u - v} v_t.$$

Furthermore, if we set $h(x, t) = e^{-\gamma t}w$, where γ is chosen later, then $h(x, t)$ satisfies the following equation:

$$\beta'(v)h_t - h_{xx} + \{\gamma\beta' + C\}h = -\{g'(u) + g'(v)\}e^{-\gamma t}u_x + g'(v)h_x. \tag{2.11}$$

Since $\beta'(v) > 0$ and $C < \infty$ for each $t \in [0, T']$ and $x \in [-n, n]$, if γ is large enough, then

$$\gamma\beta'(v) + C > 0.$$

Further, we note that

$$\begin{aligned} &h(a, t) = 0, \\ &h(-n, t) = e^{-\gamma t}\{u(-n, t) - v\} = e^{-\gamma t}\{1/n - v\} \le 0 \quad \text{for } t > 0, \\ &h(x, 0) = \varphi_n - \sigma_a\varphi_n \le 0 \quad \text{for } x \in [-n, a]. \end{aligned} \tag{2.12}$$

We shall claim that $h \le 0$ in $[-n, a] \times [0, T']$. Indeed, otherwise we take the positive maximum value of $h(x, t)$ at some point $(\bar{x}, \bar{t})$ in $(-n, a) \times (0, T']$. Then we have

$$h(\bar{x}, \bar{t}) > 0, h_x(\bar{x}, \bar{t}) = 0, h_t(\bar{x}, \bar{t}) \ge 0, \text{ and } h_{xx}(\bar{x}, \bar{t}) \le 0, \tag{2.13}$$

and for $u(x, t)$, we also have

$$u(\bar{x}, \bar{t}) > v(\bar{x}, \bar{t}), \tag{2.14}$$

$$u_x(\bar{x}, \bar{t}) = v_x(\bar{x}, \bar{t}). \tag{2.15}$$

Suppose $\bar{x} \le \eta_n(\bar{t})$. Then $u_x(\bar{x}, \bar{t}) \ge 0$. Noting this and (2.13), we see

$$\beta'(v)h_t - h_{xx} + \{\gamma\beta'(v) + C\}h > -\{g'(u) + g'(v)\}e^{-\gamma t}u_x + g'(v)h_x$$

at $(x,t)=(\bar{x},\bar{t})$. This contradicts (2.11).

Next suppose $\eta_n(\bar{t})<\bar{x}$. Then $u_x(\bar{x},\bar{t})<0$ and $v_x(\bar{x},\bar{t})>0$. These results contradict (2.15).

Hence we obtain $h\le 0$ in $[-n,a]\times[0,T')$, that is,

$$u_n(x,t)\le \sigma_a u_n(x,t) \quad \text{for } (x,t)\in[-n,a]\times\{0,T'\}. \tag{2.16}$$

Therefore,

$$u_{n,x}(a,t)\ge 0 \quad \text{for } t\in[0,T']. \tag{2.17}$$

Since $a\in[-n,a_1+\delta]$ is chosen arbitrarily, we get

$$\eta_n(t)\ge -a_1+\delta.$$

This shows (2.5). The proof is complete. □

Proof of Proposition 2.1. Using Lemmas 2.4 and 2.7 and the limit procedure of an approximate solution u_n, we can prove Proposition 2.1. Indeed, by Theorem 1.7 there exist continuous functions $\xi_i(t):[0,T)\to\mathbb{R}$ such that

$$\xi_1(0)=-a_1, \qquad \xi_2(0)=a_1 \quad \text{for } t\in[0,T) \tag{2.18}$$

and

$$\{x;u(x,t)>0\}=(\xi_1(t),\xi_2(t)) \quad \text{for each } t\in(0,T). \tag{2.19}$$

Hence, for each $t_1\in(0,T)$, there exist sequences $\{x_j^{\pm}\}$ and $\{\delta_j\}$ $(\delta_j>0)$ such that

$$x_j^-\to\xi_1(t_1) \quad \text{and} \quad x_j^+\to\xi_2(t_2) \quad \text{as } j\to\infty, \tag{2.20}$$

$$\xi_1(t)<x_j^-<x_j^+<\xi_2(t) \quad \text{for } t\in(t_1-\delta_j,t_1+\delta_j) \tag{2.21}$$

and

$$\pm u_x(x_j^{\pm},t)<0 \quad \text{for each } t\in(t_1-\delta_j,t_1+\delta_j). \tag{2.22}$$

Now we shall show that the nodal number of $u_x(\cdot,t_1)$ on $[x_j^-,x_j^+]$ is 1, namely,

$$\nu_{[x_j^-,x_j^+]}(u_x(\cdot,t_1))=1. \tag{2.23}$$

Applying Lemma 2.7 to $u_x(\cdot,t_1)$ in $[x_j^-,x_j^+]$, we can see that $\nu_{[x_j^-,x_j^+]}(u_x(\cdot,t))$ is finite for each $t\in(t_1-\delta_j,t_1+\delta_j)$ and is nonincreasing in $t\in(t_1-\delta_j,t_1+\delta_j)$, and we see that $u_x(\cdot,t_2)\in\Sigma([x_j^-,x_j^+])$ for some $t_2\in(t_1-\delta_j,t_1)$. Then we get that if n is large enough,

$$\{\eta_n(t)\}\subset(x_j^-,x_j^+) \quad \text{for } t\in(t_1-\delta_j,t_1+\delta_j). \tag{2.24}$$

In fact, assume that there exists a subsequence $\{\eta_{n_k}(t)\}\subset\{\eta_n(t)\}$ such that $x_j^+\le\eta_{n_k}(t)$. Then, by Lemma 2.4, we obtain that $u_{n_k}(x,t)$ is increasing in $x\in[x_j^-,x_j^+]$. Therefore, since $u_{n_k}(x,t)$ converges to $u(x,t)$ as $n_k\to\infty$ by Lemma 2.2, we can see

that $u(x,t)$ has the same property as $u_{n_k}(x,t)$ and so $u_x(x_j^+,t) \geq 0$. This contradicts (2.22). On the other hand assume that there exists a subsequence $\{\eta_{n_k}(t)\} \subset \{\eta_n(t)\}$ such that $\eta_{n_k}(t) \leq x_j^-$; we can also show the same contradiction.

For each $t \in (t_1 - \delta_j, t_1 + \delta_j)$ let $\eta_0(t)$ be an accumulating point of $\{\eta_n(t)\}$. Then, since $u(\cdot,t)$ is decreasing in $x \in [x_j, \eta_0(t)]$ and $u(\cdot,t)$ is nonincreasing in $x \in [\eta_0(t), x_j^+]$, namely, $u_x(x,t) \geq 0$ in $x \in [x_j^-, \eta_0(t)]$ and $u_x(x,t) \leq 0$ in $x \in [\eta_0(t), x_j^+]$, we obtain $u_x(\eta_0(t),t) = 0$ and $\eta_0(t) \in (x_j^-, x_j^+)$.

Take $t = t_2$ and assume that there is a point $x_1 \in (x_j^-, x_j)$ but η_0 such that $u_x(x_1,t_2) = 0$. It follows from $u_x(\cdot,t_2) \in \Sigma([x_j^-, x_j^+])$ that x_1 is a simple zero point of $u_x(x,t_2)$ in $[x_j^-, x_j^+]$, namely, $u_{xx}(x_1,t_2) \neq 0$. This contradicts the fact that $u_x(x,t_2) \geq 0$ (or ≤ 0) in the neighborhoods of x_1. Hence the zero points of $u_x(x,t_2)$ in $[x_j^-, x_j^+]$ coincide with $\eta_0(t_2)$, that is, $\nu_{[x_j^-, x_j^+]}(u_x(\cdot,t_2)) = 1$.

Since the nodal number of $u_x(\cdot,t)$ in $[x_j^-, x_j^+]$ is nonincreasing in t, we get (2.23). Furthermore, noting that the accumulating point is only $\eta_0(t_1)$, we see that $\eta_n(t_1) \to \eta_0(t_1)$ as $n \to \infty$.

Therefore, if $j \to \infty$ in (2.23), then we get

$$\nu_{(\xi_1(t_1), \xi_2(t_2))}(u_x(\cdot,t_1)) = 1$$

and

$$\{x \in (\xi_1(t_1), \xi_2(t_1)); u_x(x,t) = 0\} = \{\eta_0(t_1)\}.$$

Noting that $t_1 \in (0,T)$ is chosen arbitrarily and $u_x(\cdot,t_1) \in \Sigma(\xi_1(t_1), \xi_2(t_1))$ by Lemma 2.7 (ii), and setting $\eta(t) = \eta_0(t)$, we have that $\{x \in (\xi_1(t), \xi_2(t)); u_x(x,t) = 0\} = \{\eta(t)\}$ for each $t \in (0,T)$, $\eta(t)$ is C^1-function, and

$$\eta_n(t) \to \eta(t) \ (n \to \infty) \quad \text{for each } t \in (0,T). \tag{2.25}$$

(2.5) is reduced by (2.2) easily. The proof is complete. □

3. Single-point blow-up. In this section we assume (A1)–(A6) and (A9), we show the existence of single-point blow-up solutions of (0.1), (0.2), and we study the asymptotic behavior of interfaces of the blow-up solutions. To accomplish this, we also need (A7) and (A8) and the following assumptions:

(A10) There exists a C^2-function $F(v)$ such that

(i) $F(v), F'(v), F''(v) > 0$ for $v > 0$,

(ii) $\int_1^\infty d\xi/F(\xi) < \infty$, and

(iii) there are constants $c > 0$ and $v_0 > 0$ such that

$$f'F - F'f - \tfrac{1}{2}(g')^2 F \geq c(F^2 g'' + F'F) \quad \text{for } v \geq v_0.$$

Condition (A10) shows that $f(u)$ grows more rapidly than $g(u)$ and u.

Remark 3.1. (A10) is satisfied by (0.4) if $p > \max\{m, 2q - m\}$. In this case we can put $F(\xi) = \xi^{\tilde{p}/m}$ for $m < \tilde{p} < \min\{p, p + m - q\}$.

Condition (A8) is required to ensure that $u(x,t)$ is increasing in t for each $x \in \mathbb{R}$. Namely, we have the following lemma.

LEMMA 3.2. *Assume* (A1)–(A3) *and* (A8). *Let* $u(x,t)$ *be a weak solution of* (0.1), (0.2) *in* $\mathbb{R} \times (0,T)$. *Then* $u(x,t)$ *is nondecreasing in* t. *If* $u(x_0,t_0) > 0$ *for some* $(x_0,t_0) \in \mathbb{R} \times (0,T)$, *then* $\partial_t u(x,t) \geq 0$ *in the neighborhood of* (x_0,t_0).

Further, if we add (A7) and (A10), then we can get the following lemma.

LEMMA 3.3 (cf. Chen and Matano [2] and Suzuki [16]). *Assume* (A1)–(A3), (A7), (A8), *and* (A10). *Let* $\Omega = (a,b)$ *be a bounded open interval and let* $u(x,t)$ *be a positive weak solution of* (0.1) *in* $Q_T = \Omega \times (0,T)$. *Furthermore, suppose that*

$$u_x(x,t) > 0 \,[\text{or } u_x(x,t) < 0] \quad \text{in } (x,t) \in [c-\delta, c+\delta] \times (\tau, T) \tag{3.1}$$

for some $c \in (a,b)$ *and* $\delta > 0$ *with* $(c-\delta, c+\delta) \subset (a,b)$ *and some* $\tau \in (0,T)$. *Then there are no blow-up points in* $(c-\delta, c+\delta)$.

Proof. We shall prove this lemma in the case when

$$u_x(x,t) > 0 \quad \text{in } (x,t) \in (c-\delta, c+\delta) \times (\tau, T). \tag{3.2}$$

Assume $x_0 \in (c-\delta, c+\delta)$ is a blow-up point of $u(x,t)$. Then, by (3.2) and Lemma 3.2 we soon see that

$$\lim_{t \uparrow T} u(x,t) = \infty \quad \text{for } x \in (x_0, c+\delta). \tag{3.3}$$

Choose $d \in (x_0, c+\delta)$ and set

$$J = u_x - \varepsilon \rho(x) F(u(x,t)), \qquad (x,t) \in Q = (d, c+\delta) \times (\tau, T) \tag{3.4}$$

and

$$\rho(x) = \left[\sin \frac{\pi(x-d)}{c+\delta-d}\right]^2, \tag{3.5}$$

where $\varepsilon > 0$ and $t_1 \in (\tau, T)$ is chosen later. We compute

$$\begin{aligned}(\beta' J)_t - J_{xx} = {} & \varepsilon \rho A(x,t) + B(x,t) J - g' J_x \\ & - \varepsilon \beta' F u_t + \varepsilon \rho F''(u_x)^2,\end{aligned} \tag{3.6}$$

where

$$A(x,t) = f'F - F'f + \left\{-\frac{\rho'}{\rho} g' + \frac{\rho''}{\rho}\right\} F - \varepsilon\{\rho F^2 g'' + 2\rho' F' F\}$$

and

$$B(x,t) = f' + \varepsilon \rho F' g' + 2\varepsilon \rho' F' - \varepsilon g' \rho F' - g'' J - 2\varepsilon g'' \rho F.$$

Here we used the relations

$$u_{xx} = J_x + \varepsilon \rho' F + \varepsilon \rho F' J + \varepsilon^2 \rho^2 F' F$$

and

$$(u_x)^2 = J^2 + 2\varepsilon \rho F J + \varepsilon^2 \rho^2 F^2.$$

If we note that

$$\rho' = 2\lambda \sin \lambda(x-d) \cdot \cos \lambda(x-d)$$

and

$$\rho'' = 2\lambda^2 (1 - 2\rho) \quad \text{where } \lambda = \frac{\pi}{c+\delta-d},$$

then we get

$$-\frac{\rho'}{\rho}g' + \frac{\rho''}{\rho} = \frac{-2g' \sin\lambda(x-d)\cdot\cos\lambda(x-d) + 2\lambda^2(1-2\rho)}{\{\sin\lambda(x-d)\}^2}$$
$$= -4\lambda^2 + \frac{2\lambda\{\lambda - \cos\lambda(x-d)\cdot\sin\lambda(x-d)\cdot g'\}}{\{\sin\lambda(x-d)\}^2}$$
$$\geq -4\lambda^2 + \frac{2\lambda\{\lambda - \sin\lambda(x-d)\cdot|g'|\}}{\{\sin\lambda(x-d)\}^2}.$$

Hence, putting $\theta = \sin\lambda(x-d)$, we get:

$$-\frac{\rho'}{\rho}g' + \frac{\rho''}{\rho} \geq -4\lambda^2 + \frac{2\lambda(\lambda - \theta|g'|)}{\theta^2},$$

where $0 \leq \theta \leq 1$.

Set

$$h(\theta) = -4\lambda + \frac{2\lambda(\lambda - \theta|g'|)}{\theta^2}$$

and assume that θ is independent of g'. Then we see that $h(\theta)$ takes the minimum value

$$h\left(\frac{2\lambda}{|g'|}\right) = -4\lambda^2 - \frac{1}{2}(g')^2 \quad \text{at } \theta = \frac{2\lambda}{|g'|},$$

since $h'(\theta) = -4\lambda^2\theta^{-3} + 2\lambda\theta^{-2}g'$, and $h(\theta) = 0$ is reduced to $\theta = 2\lambda/|g'|$. Therefore, we have

$$-\frac{\rho'}{\rho}g' + \frac{\rho''}{\rho} \geq -4\lambda^2 - \frac{1}{2}(g')^2. \tag{3.7}$$

Thus, we get the lower bound estimate of $A(x,t)$:

$$A(x,t) \geq f'F - F'f + (4\lambda^2 + \tfrac{1}{2}(g')^2)F - \varepsilon\{\rho F^2 g'' + 2|\rho'|F'F\}. \tag{3.8}$$

Considering (A10), Lemma 3.2, and the fact that $F'(u)$ goes to infinity as $u \to \infty$, if t_1 is close enough to T we have

$$(\beta' J)_t - J_{xx} \geq B(x,t)J - g'J_x. \tag{3.9}$$

On the other hand,

$$J(d,t) = u_x(d,t) > 0, \qquad J(c+\delta,t) = u_x(c+\delta,t) > 0, \tag{3.10}$$

and

$$J(x,t_1) > 0 \quad \text{(by (3.2))} \quad \text{for small enough } \varepsilon > 0. \tag{3.11}$$

Applying the maximum principle to $J(x,t)$, we obtain

$$J(x,t) > 0 \quad \text{in } (t_1,T)\times(d,c+\delta),$$

namely,

$$\frac{u_x}{F} > \varepsilon\rho \quad \text{in } (t_1,T)\times(d,c+\delta). \tag{3.12}$$

Integrating this inequality over $d \leq x \leq c+\delta$ yields

$$\int_{u(d,t)}^{u(c+\delta,t)} \frac{du}{F(u)} > \varepsilon \int_d^{c+\delta} \rho(x)dx \quad \text{in } t_1 < t < T. \tag{3.13}$$

The right-hand side of (3.13) is a positive constant, while the left-hand side tends to zero as $t \uparrow T$ by virtue of condition (A10)(ii) and equation (3.3). This contradiction shows that x_0 is not a blow-up point of $u(x,t)$. The proof is complete. □

THEOREM 3.4. *Let $u(x,t)$ and $\xi_i(t)$ be as in Theorem* 1.7, *and let S be the blow-up set of $u(x,t)$. Furthermore, assume* (A7)–(A10). *Then*

$$S = \{\eta_0\} \tag{3.14}$$

for some $\eta_0 \in [-a_1+\delta, \infty]$ with a small $\delta > 0$ and

$$-\infty < \lim_{t\uparrow T} \xi_1(t). \tag{3.15}$$

Proof (see Friedman and Lacey [4]). By Proposition 2.1, there exists a C^1-function $\eta(t) : (0,T) \to \mathbb{R}$ such that

$$\{x \in (\xi_1(t), \xi_2(t)); u_x(x,t) = 0\} = \{\eta(t)\} \tag{3.16}$$

for each $t \in (0,T)$ and

$$-a_1 + \delta \leq \eta(t) \quad \text{for all } t \in (0,T), \tag{3.17}$$

where $\xi_i(t)$ is defined by (1.11), (1.12). Therefore, we see that

$$u_x(x,t) > 0 \quad \text{for } \xi_1(t) < x < -a_1+\delta, \quad 0 < t < T, \tag{3.18}$$

and it follows from Lemma 3.3 that

$$\{x; x < -a_1+\delta\} \subset S^c, \tag{3.19}$$

where S is the blow-up set of $u(x,t)$.

Here, if we show that

$$\lim_{t\uparrow T} \eta(t) = \eta_0 \tag{3.20}$$

exists, then by Lemmas 3.2 and 3.3 we can obtain the results of Theorem 3.4. Hence, we shall prove (3.20).

Assume that $\lim_{t\uparrow T} \eta(t)$ does not exist. Then, if we set $\eta_- = \liminf_{t\uparrow T} \eta(t)$ and $\eta_+ = \liminf_{t\uparrow T} \eta(t)$,

$$-a_1 + \delta \leq \eta_- < \eta_+ \leq \infty. \tag{3.21}$$

Choose $-a_1+\delta < s_1 < \eta_-$ and $\eta_- < s_2 < \eta_+$ such that

$$\alpha = (s_1+s_2)/2 \in (\eta_-, \eta_+). \tag{3.22}$$

Then, since $\lim_{t\uparrow T} u(\eta(t),t) = \infty$, by Lemma 3.2 we get $\lim_{t\uparrow T} u(x,t) = \infty$ for each $x \in (\eta_-, \eta_+)$. Hence, if T_0 is chosen close enough to T, we obtain

$$u_x(x,T_0) > 0 \quad \text{for } s_1 < x < s_2 \tag{3.23}$$

and

$$u(s_1,t) < u(s_2,t) \quad \text{for } t \in (T_0,T). \tag{3.24}$$

Set $w = u - v$ where $v(x,t) = \sigma_\alpha u = u(2\alpha - x,t)$, and consider $w(x,t)$ in the rectangle region $R = \{s_1 < x < \alpha, T_0 < t < T\}$. Then, we see

$$w(x,T_0) = u(x,T_0) - \sigma_\alpha u(x,T_0) \le 0 \quad \text{in } [s_1,\alpha] \tag{3.25}$$

and

$$w(x,s_1) = u(s_1,t) - u(s_2,t) \le 0 \quad \text{in } t \in [T_0,T). \tag{3.26}$$

By the same method as we used in the proof of (2.16) we obtain

$$w(x,t) \le 0 \quad \text{for } (x,t) \in [s_1,\alpha] \times [T_0,T). \tag{3.27}$$

Since $w(\alpha,t) = 0$, we get

$$\tfrac{1}{2} w_x(a,t) = u_x(\alpha,t) \ge 0 \quad \text{for } t \in [T_0,T). \tag{3.28}$$

This is a contradiction of $\alpha \in (\eta_-, \eta_+)$. Therefore, we obtain (3.20) and $S = \{\eta_0\}$. The proof is complete. □

4. The upper bound estimates and bounded-point blow-up. In this section we show $\eta_0 < \infty$ where η_0 is as in Theorem 3.4. In order to show this, we need the upper bound estimates of the blow-up solution of (0.1), (0.2). We must further assume (A11) on $f(u)$ such that $f(u)$ grows more rapidly than u and $g(u)$, and we choose the suitable initial data $\varphi(x)$ corresponding to (A11).

(A11) There exists a C^2-function $\Phi(v)$ such that

(i) $\Phi, \Phi', \Phi'' > 0$ for $v > 0, \Phi(0) = \Phi'(0) = 0$ and $f(v)/\beta'(v) \ge \Phi(v)$ near $v = 0$;

(ii) $\int_1^\infty d\xi/\Phi(\xi) < \infty$;

(iii) There are constants $C > 0$ and $v_1 > 0$ such that

$$4\Phi''(f'\Phi - \Phi' f) \ge (g'')^2\Phi \quad \text{for } v \ge v_1$$

and

$$\frac{(g'')^2\Phi}{4\Phi''\beta'} + \frac{f\Phi'}{\Phi\beta'} < C \quad \text{for } 0 \le v \le v_1;$$

(iv) $$\int_0^1 \sup_{0 \le v \le H^{-1}(t)} \frac{g'(v)}{\beta'(v)}\, dt < \infty,$$

where $H(\xi) = \int_\xi^\infty \frac{d\eta}{\Phi(\eta)}$.

$$\varphi'' - g(\varphi)' + f(\varphi) \ge \Phi(\varphi) \quad \text{in } \mathcal{D}', \tag{A12}$$

where $\Phi(\eta)$ is as in (A11).

Remark 4.1. (A11) is satisfied by (0.4) if $p > \max\{m, 2q-1, m+q-1\}$ and $2q \geq m+1$. In this case, we can put $\Phi(\xi) \approx \xi^{p/m-\delta}$ for sufficiently small $\delta > 0$ (as $\xi \to \infty$) and $\Phi(\xi) = m\xi^{(p+m-1)/m}$ (as $\xi \to 0$). In general, we can not replace $\Phi(\xi)$ by the function $F(\xi)$ in (A10).

LEMMA 4.2. *Assume* (A1)–(A12). *Let* $u_n(x,t)$ *be a solution of the regularized problem*

$$\begin{aligned} &\partial_t \beta(u) - u_{xx} + g(u)_x = f(u), && (x,t) \in [-n,n] \times (0,T), \\ &u(x,0) = \varphi_n(x), && x \in [-n,n] \times (0,T), \\ &u(\pm n,t) = \Phi(1/n)t + 1/n, && t > 0, \end{aligned} \tag{4.1}$$

with the blow-up time T_n. *Let* $\varphi_n(x)$ *be as in Lemma* 2.2, *and let us add the following condition to it*:

$$\varphi_n'' - \{g(\varphi_n)\}' + f(\varphi_n) \geq \Phi(\varphi_n) \quad \text{in } \mathcal{D}' \tag{4.2}$$

(*the existence of the above* $\varphi_n(x)$ *is guaranteed by assumption* (A12)). *Then, for some* $c_1 > 0$,

$$u_n(x,t) \leq H^{-1}(c_1(T_n' - t)) \quad \text{for } (x,t) \in \mathbb{R} \times [0, T_n'), \tag{4.3}$$

where $H(\xi) = \int_\xi^\infty d\eta/\Phi(\eta)$ *and* $T_n' = \min\{T_n, T\}$.

LEMMA 4.3. *Lemmas* 2.2 *and* 2.4 *hold for large* n *if* (2.3) *is replaced by* (4.1). *Furthermore, it holds that* $u_n(x,t) \geq 1/n + \Phi(1/n)t$ *for* $(x,t) \in [-n,n] \times [0, T_n)$.

Proof. Noting the condition (A11)(i) and using the comparison principle, we obtain $u_n(x,t) \geq 1/n + \Phi(1/n)t$. Proceeding with a proof similar to that of Lemma 2.4, we get the assertion of Lemma 4.3. □

Proof of Lemma 4.2 (see Friedman and Lacey [4]). Set

$$J = u_t - c(t)\Phi(u) \tag{4.4}$$

where $u = u_n$, and

$$c(t) = \frac{\Phi(1/n)}{\Phi(1/n + T\Phi(1/n))} e^{-Ct}$$

where $C > 0$ is as in (A11)(iii). We compute $(\beta' J)_t - J_{xx}$ in the following way:

$$(\beta' J)_t - J_{xx} = B(x,t)J - g' J_x + c(t)A(x,t),$$

where

$$A(x,t) = f'\Phi - \Phi' f - c\beta''\Phi^2 + C\beta'\Phi - g''\Phi u_x + \Phi''(u_x)^2$$

and

$$B(x,t) = -g'' u_x - c\beta''\Phi + f'.$$

Here we used the next relation:

$$u_{xt} = J_x + c\Phi' u_x.$$

We further compute that

$$\begin{aligned} A(x,t) &= \Phi''\left\{(u_x)^2 - \frac{g''}{\Phi''}\Phi u_x\right\} + f'\Phi - \Phi' f - c\beta''\Phi^2 + C\beta'\Phi \\ &= \Phi''\left(u_x - \frac{g''}{2\Phi''}\Phi\right)^2 - \frac{(g'')^2\Phi^2}{4\Phi''} + f'\Phi - \Phi' f - c\beta''\Phi^2 + C\beta'\Phi \\ &\geq \frac{4\Phi''[f'\Phi - \Phi' f - c\beta''\Phi^2 + C\beta'\Phi] - (g'')^2\Phi^2}{4\Phi''}. \end{aligned}$$

Noting $\beta' > 0$ and condition (A11) we get

$$4C\Phi''\Phi\beta' - 4\Phi''\Phi' f - (g'')\Phi^2 \geq 4\beta'\Phi''\Phi\left\{C - \frac{\Phi' f}{\Phi\beta'} - \frac{(g'')^2\Phi}{4\Phi''\beta'}\right\} \geq 0 \quad \text{for } 0 \leq v \leq v_1.$$

Hence, considering condition (A11) again, and noting $\beta'' \leq 0$, we see that

$$A(x,t) \geq 0. \tag{4.5}$$

Thus, we have

$$(\beta' J)_t - J_{xx} \geq B(x,t)J - g' J_x. \tag{4.6}$$

On the other hand,

$$\begin{aligned} J(\pm n,t) &= u_t(\pm n,t) - c(t)\Phi(u_n(\pm n,t)) \\ &= \Phi(1/n) - \frac{\Phi(1/n)}{\Phi(1/n + T\Phi(1/n))} e^{-Ct}\Phi(1/n + \Phi(1/n)t) \geq 0 \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} J(x,0) &= u_{n,t}(x,0) - \Phi(u_n(x,0)) \\ &= \varphi'' - \{g(\varphi_n)\}' + f(\varphi_n) - \Phi(\varphi_n) \geq 0 \quad \text{(by (4.2))}. \end{aligned} \tag{4.8}$$

Applying the maximum principle to J, we obtain

$$J(x,t) \geq 0 \quad \text{in } [-n,n] \times [0, T_n'), \tag{4.9}$$

that is,

$$u_{n,t} \geq c(t)\Phi(u_n).$$

Here we note that

$$\begin{aligned} \frac{\Phi(1/n)}{\Phi(1/n + T\Phi(1/n))} &= \frac{\Phi(1/n)}{\Phi(1/n) + T\Phi(1/n)\int_0^1 \Phi'(1/n + T\Phi(1/n)s)\,ds} \\ &= \frac{1}{1 + T\int_0^1 \Phi'(1/n + T\Phi(1/n)s)\,ds} \geq 1/2 \quad \text{for large } n. \end{aligned}$$

Hence, if we put $c_1 = \frac{1}{2}e^{-CT}$, we get

$$u_{n,t} \geq c_1\Phi(u_n). \tag{4.10}$$

Integrate (4.10) over $[t, T_n')$ for each $x \in [-n, n]$. Then, we have

$$\int_t^{T_n'} \frac{u_{n,t}}{c_1 \Phi(u_n)} dt \geq T_n' - t. \tag{4.11}$$

Setting $H(\xi) = \int_\xi^\infty d\eta/\Phi(\eta)$, we have

$$-\frac{1}{c_1}[H(u_n)]_t^{T_n'} \geq T_n' - t,$$

that is,

$$-\frac{1}{c_1}\{H(u_n(x, T_n')) - H(u_n(x,t))\} \geq T_n' - t.$$

Therefore,

$$H(u_n(x,t)) \geq c_1(T_n' - t) \tag{4.12}$$

Since $H(\xi)$ is a decreasing function in ξ, we obtain

$$u_n(x,t) \leq H^{-1}(c_1(T_n' - t)).$$

The proof is complete. □

THEOREM 4.4. *In Theorem* 3.4, *if we further assume* (A11) *and* (A12), *we get*

$$\eta_0 < \infty \tag{4.13}$$

and

$$\lim_{t \uparrow T} \xi_2(t) < \infty. \tag{4.14}$$

Proof. Let $T' \in (0, T)$ be fixed and set $h(t) = H^{-1}(c_1(T' - t))$. Then by Lemma 4.2, if n is large enough, $T' < T_n'$ and

$$u_n(x,t) \leq H^{-1}(c_1(T_n' - t)) \leq h(t). \tag{4.15}$$

Put

$$v(x,t) = u_n(x + k(t), t) \tag{4.16}$$

where $k(t) = \int_0^t l(h(s))ds$ and $l(\xi) = \sup_{0 \leq v \leq \xi}(g'(v)/\beta'(v))$.

Then $v(x,t)$ satisfies the following equation:

$$\beta'(v)v_t - v_{xx} = \beta' \times \left\{k' - \frac{g'(v)}{\beta'(v)}\right\} v_x + f(v).$$

Since

$$\frac{g'(v(x,t))}{\beta'(v(x,t))} = l(v) \leq l(h(t)) = k'(t),$$

we obtain

$$k'(t) - \frac{g'(v)}{\beta'(v)} \geq 0 \quad \text{for } (x,t) \in [-k(t) - n, -k(t) + n] \times [0, T'). \tag{4.17}$$

On the other hand, noting condition (A11)(iv) we see that

$$
\begin{aligned}
k(t) \le k(T') &= \int_0^{T'} l(H^{-1}(c_1(T'-s)))\,ds \\
&= \frac{1}{c_1}\int_0^{c_1T'} l(H^{-1}(t))\,dt \quad (\text{put } t = c_1(T'-s)) \\
&\le k(T) = \frac{1}{c_1}\int_0^{c_1T} l(H^{-1}(t))\,dt \\
&< \infty \quad \text{for } t \in (0,T').
\end{aligned} \tag{4.18}
$$

Thus, for each $a \ge a_1$ there exists N such that

$$
a \le n - k(T) \le n - k(t) \quad \text{for all } n \ge N \quad \text{and each} \quad t \in [0,T']. \tag{4.19}
$$

Put $\tilde{v}(x,t) = \sigma_a v = v(2a-x,t)$. Then we can consider

$$
w = v(x,t) - \tilde{v}(x,t) \quad \text{in} \bigcup_{0\le t\le T'} [a, n-k(t)] \times \{t\}
$$

since $n - k(T) \le n - k(t)$ for $t \in [0,T']$. As w satisfies the equation

$$
\begin{aligned}
\beta'(v)w_t - w_{xx} + c(x,t)w = &\left\{\beta'(\tilde{v})\left(k' - \frac{g'(\tilde{v})}{\beta'(\tilde{v})}\right) + \beta'(v)\left(k' - \frac{g'(v)}{\beta'(v)}\right)\right\} v_x \\
&- \left\{\beta'(\tilde{v})\left(k' - \frac{g'(\tilde{v})}{\beta'(\tilde{v})}\right)\right\} w_x
\end{aligned} \tag{4.20}
$$

where

$$
c(x,t) = -\frac{f(v) - f(\tilde{v})}{v - \tilde{v}} + \frac{\beta'(v) - \beta'(\tilde{v})}{v - \tilde{v}}\tilde{v}_t,
$$

if we set $h(x,t) = e^{-\gamma t}w$ where γ is chosen later, then h satisfies the following equation:

$$
\begin{aligned}
&\beta'(v)h_t - h_{xx} + \{\gamma\beta'(v) + c\}h \\
&\quad = \left\{\beta'(\tilde{v})\left(k' - \frac{g'(\tilde{v})}{\beta'(\tilde{v})}\right) + \beta'(v)\left(k' - \frac{g'(v)}{\beta'(v)}\right)\right\} e^{-\gamma t}v_x \\
&\qquad - \left\{\beta'(\tilde{v})\left(k' - \frac{g(v)}{\beta'(\tilde{v})}\right)\right\} h_x.
\end{aligned} \tag{4.21}
$$

Since $\beta'(v) > 0$ and $c < \infty$, choosing γ large enough, we get

$$
\gamma\beta'(v) + c > 0. \tag{4.22}
$$

On the other hand, as $u(x,t) \ge 1/n + \Phi(1/n)t$ for $(x,t) \in [-n,n] \times [0,T')$ and $v(n - k(t),t) = u(n,t) = 1/n + \Phi(1/n)t$, we obtain

$$
\begin{aligned}
h(n-k(t),t) &= e^{-\gamma t}\{v(n-k(t),t) - \sigma_a v(n-k(t),t)\} \\
&= e^{-\gamma t}\left\{\frac{1}{n} + \Phi\left(\frac{1}{n}\right)t - \sigma_a v(n-k(t),t)\right\} \le 0
\end{aligned} \tag{4.23}
$$

and

$$h(a,t) = e^{-\gamma t}\{v(a,t) - v(a,t)\} = 0. \tag{4.24}$$

Noting the condition on φ_n in Lemma 2.2, we also obtain

$$\begin{aligned} h(x,0) &= u_n(x + k(0), 0) - \sigma_a u_n(x + k(0), 0) \\ &= \varphi_n(x) - \sigma_a \varphi_n(x) \leq 0 \quad \text{for each } x \in [a,n]. \end{aligned} \tag{4.25}$$

We claim that $h \leq 0$ in $\cup_{0 \leq t \leq T'}[a, n - k(t)] \times \{t\}$. Suppose

$$(\bar{x}, \bar{t}) \in \cup_{0 < t \leq T'}(a, n - k(t)) \times \{t\}$$

is a maximum point and $h(\bar{x}, \bar{t}) > 0$. Then

$$h_t(\bar{x}, \bar{t}) \geq 0, \quad h_{xx}(\bar{x}, \bar{t}) \leq 0, \quad h_x(\bar{x}, \bar{t}) = 0, \tag{4.26}$$

namely,

$$v(\bar{x}, \bar{t}) > \tilde{v}(\bar{x}, \bar{t}), \tag{4.27}$$

$$v_x(\bar{x}, \bar{t}) = \tilde{v}_x(\bar{x}, \bar{t}). \tag{4.28}$$

Assume $\bar{x} \geq \eta_n(\bar{t}) - k(\bar{t})$. Then

$$v_x(\bar{x}, \bar{t}) \leq 0.$$

Noting this, (4.17), and (4.26), we see that

$$\begin{aligned} &\beta'(v)h_t - h_{xx} + \{\gamma\beta'(v) + c\}h \\ &> \left\{\beta'(\tilde{v})\left(k' - \frac{g'(\tilde{v})}{\beta'(\tilde{v})}\right) + \beta'(v)\left(k' - \frac{g'(v)}{\beta'(v)}\right)\right\} e^{-\gamma t} v_x \\ &\quad - \left\{\beta'(\tilde{v})\left(k' - \frac{g(v)}{\beta'(\tilde{v})}\right)\right\} h_x. \end{aligned}$$

This contradicts (4.21).

On the other hand, assume $\bar{x} < \eta_n(\bar{t}) - k(\bar{t})$. Then $v_x(\bar{x}, \bar{t}) > 0$ and $v_x(\bar{x}, \bar{t}) < 0$. These results contradict (4.28). Hence, we obtain $h \leq 0$, that is,

$$w(x,t) \leq 0, \qquad (x,t) \in [a, n - k(t)) \times [0, T'], \tag{4.29}$$

or

$$u_n(x + k(t), t) \leq u_n(2a - x + k(t), t), \qquad (x,t) \in [a, n - k(t)] \times [0, T']. \tag{4.30}$$

Here, if $n \to \infty$, then

$$u(x + k(t), t) \leq u(2a - x + k(t), t) \quad \text{for } (x,t) \in [a, \infty) \times [0, T']. \tag{4.31}$$

Putting $x' = x + k(t), a' = a + k(t)$, and noting $k(t) \leq k(T)$, we obtain for each $a' \in [a_1 + k(T), \infty)$

$$u(x', t) \leq \sigma_{a'} u(x', t) \quad \text{for } (x', t) \in [0, T'].$$

Since this shows that $u(x, t)$ is a decreasing function in $x \in [a_1 + k(T), \infty)$ for $t \in [0, T']$, we get

$$\eta(t) \leq a_1 + k(T).$$

As $T' \in (0, T)$ is chosen arbitrarily, we conclude that

$$\eta(t) \leq a_1 + k(T) \quad \text{for } t \in [0, T). \tag{4.32}$$

Hence, we get

$$\eta_0 = \lim_{t \uparrow T} \eta(t) \leq a_1 + k(T),$$

and noting Theorem 3.4, we obtain

$$S = \{\eta_0\}.$$

Therefore, by virtue of Lemma 1.6, we also obtain (4.14). The proof is complete. □

REFERENCES

[1] S. ANGENENT, *The zeroset of a solution of a parabolic equation*, J. Reine Angew. Math., 390 (1988), pp. 79–96.

[2] X.-Y. CHEN AND H. MATANO, *Convergence, asymptotic periodicity and finite-point blow-up in one-dimensional semilinear heat equations*, J. Differential Equations, 78 (1819), pp. 160–190.

[3] J. I. DIAZ AND R. KERSNER, *On a nonlinear degenerate parabolic equation in infiltration or evaporation through a porous medium*, J. Differential Equations, 69 (1987), pp. 368–403.

[4] A. FRIEDMAN AND A. A. LACEY, *Blowup of solutions of semilinear parabolic equations*, J. Math. Anal. Appl., 132 (1988), pp. 171–186.

[5] A. FRIEDMAN AND B. MCLEOD, *Blow-up of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425–447.

[6] V. A. GALAKTIONOV, *Proof of the localization of unbounded solutions of the non-linear partial differential equation* $u_t = (u^\sigma u_x)_x + u^\beta$, Differential'nye Uravneniya, 21 (1985), pp. 15–23; Differential Equations, 21 (1985), pp. 11–18.

[7] Y. GIGA AND R. V. KOHN, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297–319.

[8] B. H. GILDING, *A nonlinear degenerate parabolic equation*, Ann. Scuola Norm. Sup. Pisa, 4 (1977), pp. 393–432.

[9] T. IMAI AND K. MOCHIZUKI, *On blow-up solutions for quasilinear degenerate parabolic equations*, Publ. Res. Inst. Math. Sci., 27 (1991), pp. 695–709.

[10] T. IMAI, K. MOCHIZUKI, AND R. SUZUKI, *On blow-up sets for the parabolic equation* $\partial_t \beta(u) = \Delta u + f(u)$ *in a ball*, J. Fac. Sci. Shinshu Univ., 25 (1990), pp. 51–58.

[11] B. F. KNERR, *The porous medium equation in one dimension*, Trans. Amer. Math. Soc., 234 (1977), pp. 381–415.

[12] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Math. Monographs 23, American Mathematical Society, Providence, RI, 1968.

[13] H. MATANO, *Nonincreasing of the lap-number of a solution for a one-dimensional semilinear parabolic equation*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 29 (1982), pp. 401–441.

[14] K. MOCHIZUKI AND R. SUZUKI, *Blow-up sets and asymptotic behavior of interfaces for quasilinear degenerate parabolic equations in* $\mathbb{R}^n$, J. Math. Soc. Japan, 44 (1992), pp. 485–504.

[15] O. A. OLEINIK, A. S. KALASHNIKOV, AND CHZOU YUI-LIN, *The Cauchy problem and boundary problems for equations of the type of nonstationary filtration*, Izv. Akad. Nauk SSSR Ser. Mat., 22 (1958), pp. 667–704. (In Russian.)

[16] R. SUZUKI, *On blow-up sets and asymptotic behavior of one dimensional quasilinear degenerate parabolic equations*, Publ. Res. Inst. Math. Sci., 27 (1991), pp. 375–398.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 98–111, January 1995

DYNAMICAL BEHAVIOR OF SOLUTIONS OF A SEMILINEAR HEAT EQUATION WITH NONLOCAL SINGULARITY *

KENG DENG†

Abstract. The heat equation with a nonlocal nonlinearity $u_t = u_{xx} + \varepsilon\|u(\cdot,t)\|^q/(1-u), 0 < x < 1, \varepsilon, q > 0$, subject to $u(0,t) = u(1,t) = 0$ is studied. Stability-instability is analyzed and finite time quenching results are given. Discussions are also extended to more general problems.

Key words. nonlocal parabolic equation, equilibrium state, stability, quenching, global existence

AMS subject classifications. 35K05, 35K20, 35K55, 35K57, 35K60

1. Introduction.

In this paper, the following initial-boundary value problem is considered:

$$\text{(D)}\qquad \begin{aligned} u_t &= u_{xx} + \varepsilon\|u(\cdot,t)\|^q/(1-u), && 0 < x < 1, \quad t > 0,\\ u(0,t) &= u(1,t) = 0, && t > 0,\\ u(x,0) &= u_0(x), && 0 \le x \le 1. \end{aligned}$$

Here $\varepsilon, q > 0$ and $u_0(x)$ is a continuous function with $u_0(0) = u_0(1) = 0$. Moreover, we take

$$\|u(\cdot,t)\| = \int_0^1 |u(x,t)|\,dx$$

and require that $0 \le u_0 < 1$. Then, by the maximum principle, $u(\cdot,t) \ge 0$ for all t in the existence interval.

There are two reasons for considering this problem. One comes from the physical motivation, since the above problem is closely related to a popular model arising in the study of a polarization phenomenon in ionic conductors as follows (in our notation):

$$\text{(K)}\qquad \begin{aligned} u_t &= u_{xx} + \varepsilon/(1-u), && 0 < x < 1, \quad t > 0,\\ u(0,t) &= u(1,t) = 0, && t > 0,\\ u(x,0) &= 0, && 0 \le x \le 1. \end{aligned}$$

In [5], the following were proved for (K):

(a) If $\varepsilon > 8$, there is a finite time T such that $\lim_{t\to T^-} u(\frac{1}{2},t) = 1$;

(b) Whenever (a) holds, $\lim_{t\to T^-} \max_x u_t(x,t) = +\infty$.

This kind of phenomenon is known as quenching, that is, the solution of the equation remains bounded, whereas its derivatives blow up at some moment. It is easily seen that (K) is a special case of (D) (when $q = 0, u_0(x) \equiv 0$). Furthermore, the nonlocal term $\int_0^1 |u(x,t)|\,dx$ can be treated as an average value of $u(x,t)$ on $[0,1]$, which means that we intend to use all the information over the whole interval. Thus, as many other nonlocal mathematical models have been formulated from physical phenomena over

* Received by the editors December 24, 1991; accepted for publication (in revised form) May 12, 1993.

† Department of Mathematics, University of Southwestern Louisiana, Lafayette, Louisiana 70504-1010.

the last few years, a more precise description of the "real world" may be expected. Actually, we shall see that the nonlocal nonlinearity has a remarkable effect on the dynamical behavior of solutions.

The other reason for investigating problem (D) comes from a purely mathematical point of view. Since 1975, when the results for (K) appeared, problem (K) and its various generalizations have been extensively studied (see [6], [7], and the literature cited therein). However, to the best of our knowledge, no one has done any problem of the type described here. Due to the presence of the nonlocal term $\|u(\cdot,t)\|^q$ in the equation, the discussion becomes more complicated, and certain conventional arguments may not apply. Therefore, we want to undertake a study.

The plan of our paper is as follows: in the next section, we present the comparison theorem and local existence of solutions for (D). In §3, we characterize the set of stationary solutions of (D). In §4, we establish stability and quenching results for (D). Finally, we discuss more general problems.

2. Comparison and local existence. For simplicity, let $D_T = (0,1) \times (0,T)$ and $D_T \cup \Gamma_T = [0,1] \times [0,T)$. We begin with the definitions of subsolution and supersolution of (D).

DEFINITION. *A function $u(x,t)$ is called a subsolution of* (D) *on D_T if $u \in C^{2,1}(D_T) \cap C(D_T \cup \Gamma_T)$ satisfies*

$$\text{(C)} \qquad \begin{aligned} u_t &\le u_{xx} + \varepsilon\|u(\cdot,t)\|^q/(1-u), && 0<x<1, \quad 0<t<T, \\ u(x,t) &\le 0, && x=0,1, \quad 0<t<T, \\ u(x,0) &\le u_0(x), && 0 \le x \le 1. \end{aligned}$$

A supersolution is defined by (C) *with reversed inequalities.*

In the sequel, we shall use the comparison principle to conduct our discussion. For this reason, we state the following theorem.

THEOREM 2.1. *Let $u(<1)$ be a subsolution and $v(<1)$ be a supersolution of problem* (D). *Then $u \le v$ on D_T.*

Proof. Suppose $u, v \le \delta < 1$ on D_T. Then $f(s) = 1/(1-s)$ is continuously differentiable for $0 \le s \le \delta$, and the conclusion follows from the relevant theorem in [3].

COROLLARY 2.2. *If $u_0'' + \varepsilon\|u_0\|^q/(1-u_0) \ge 0 (\le 0)$ on $(0,1)$, then $u_t(x,t) \ge 0 (\le 0)$ on D_T.*

Proof. The condition on u_0 implies that u_0 is a subsolution (supersolution) of (D). Thus $u(x,t) \ge u_0(x) (\le u_0(x))$ on D_T. Then let $v(x,t) = u(x,t+h) (0 < h < \frac{T}{2})$. In D_{T-h}, we find that $v(x,0) = u(x,h) \ge u_0(x) (\le u_0(x))$, and $v(x,t)$ is a solution, and thus a supersolution (subsolution), of (D). It follows that $u(x,t+h) \ge u(x,t)$ $(\le u(x,t))$ for $h > 0$ arbitrarily small, and hence $u_t \ge 0 (\le 0)$.

By means of the comparison theorem, we now establish the local existence of solutions of (D).

THEOREM 2.3. *For nontrivial initial datum, there exists a $T_0 < \infty$ such that problem* (D) *has a unique nonnegative solution on D_{T_0}.*

Proof. Clearly $u \equiv 0$ is a subsolution. If there is a supersolution v with $v < 1$ on D_{T_0} for some $T_0 < \infty$, then by the comparison, any solution $u(x,t)$ should be bounded away from one. Recalling the local existence theorem in [3], we then obtain the desired result.

Note that if $v < 1$ then $\|v(\cdot,t)\|^q < 1$ on $0 < t < T_0$ for any $q > 0$. Thus it suffices to find a solution of the ordinary differential equation

$$\text{(O)} \qquad \frac{dv}{dt} = \frac{\varepsilon}{1-v}, \qquad v(0) = \max_{0 \le x \le 1} u_0(x).$$

Problem (O) has a solution of the form $v = 1 - [(1 - v(0))^2 - 2\varepsilon t]^{1/2}$. For sufficiently small t, v remains bounded by one and hence v is an apppropriate supersolution.

3. The stationary solutions. For the stationary solutions of (D), we need to solve

$$\text{(S)} \qquad \begin{aligned} &v'' + \varepsilon \|v\|^q/(1-v) = 0, \qquad 0 < x < 1, \\ &v(0) = v(1) = 0. \end{aligned}$$

If $v(x)$ is a nonnegative classical solution of (S), then by the strong maximum principle, $v(x) > 0$, and consequently $v'' < 0$ on $(0,1)$. Thus v has exactly one maximum at $\xi \in (0,1)$. For convenience, let

$$F(s) = \log \frac{1}{1-s}.$$

Then v also solves

$$\tfrac{1}{2}(v'(x))^2 + \varepsilon \|v\|^q F(v(x)) = \varepsilon \|v\|^q F(\mu), \tag{3.1}$$

where $\mu = v(\xi)$.

Integrating (3.1), we have

$$\int_v^\mu \frac{1}{\sqrt{F(\mu) - F(\eta)}}\, d\eta = \sqrt{2\varepsilon}\|v\|^{q/2}|\xi - x|. \tag{3.2}$$

Since $v(0) = v(1) = 0$, from

$$\int_0^\mu \frac{1}{\sqrt{F(\mu) - F(\eta)}}\, d\eta = \sqrt{2\varepsilon}\|v\|^{q/2}\xi = \sqrt{2\varepsilon}\|v\|^{q/2}(1-\xi), \tag{3.3}$$

it follows that $\xi = \frac{1}{2}$.

Let

$$G(\mu) = \int_0^\mu \frac{1}{F(\mu) - F(\eta)}\, d\eta.$$

Then (3.3) is equivalent to

$$G(\mu) = \sqrt{\varepsilon/2}\|v\|^{q/2}. \tag{3.4}$$

Note that $v(1-x)$ is also a solution of (S). Combining this fact with (3.2) ensures that there is exactly one solution of (S) with $v(\frac{1}{2}) = \mu$. Thus for $0 < x \le \frac{1}{2}$, $v(x)$ is implicitly given by

$$\int_0^{v(x)} \frac{1}{\sqrt{F(\mu) - F(\eta)}}\, d\eta = \sqrt{2\varepsilon}\|v\|^{q/2}x \tag{3.5}$$

and by $v(x) = v(1-x)$ if $\frac{1}{2} < x < 1$, with μ and $\|v\|$ satisfying (3.4). Therefore, we should focus our attention on (3.4). However, for each given ε, in order to count the

number of μ, we need an additional relation between μ and ε that is independent of $\|v\|$. To this end, we let

$$y = \int_0^x (1 - v(s))\, ds, \qquad Y = \int_0^1 (1 - v(s))\, ds = 1 - \|v\|. \tag{3.6}$$

Thus, in lieu of (S), with $h(y) = (1 - v(x(y)))^2$, we find

$$\begin{aligned} &h_{yy} - 2\varepsilon \|v\|^q / h = 0, \qquad 0 < y < Y, \\ &h(0) = h(Y) = 1. \end{aligned} \tag{3.7}$$

Then, using a scale change of variable $z = y/Y$, and setting $w(z) = 1 - h(y)$, we obtain

$$\begin{aligned} &w_{zz} + 2\varepsilon \|v\|^q (1 - \|v\|)^2 / (1 - w) = 0, \qquad 0 < z < 1, \\ &w(0) = w(1) = 0. \end{aligned} \tag{3.8}$$

Note that if $\lambda = \max_{0 \le z \le 1} w(z)$, then $\lambda = 2\mu - \mu^2$ with $\mu = \max_{0 \le x \le 1} v(x)$. By a similar reasoning, we get

$$G(\lambda) = \sqrt{\varepsilon} \|v\|^{q/2} (1 - \|v\|). \tag{3.9}$$

Then the combination of (3.4) and (3.9) yields

$$\|v\| = (\sqrt{2} G(\mu) - G(\lambda)) / (\sqrt{2} G(\mu)). \tag{3.10}$$

Substituting (3.10) into (3.4) and letting

$$K(\mu) = 2 \left(\frac{\sqrt{2} (G(\mu))^{1 + (2/q)}}{\sqrt{2} G(\mu) - G(\lambda)} \right)^q,$$

we then have

$$K(\mu) = \varepsilon. \tag{3.11}$$

On the other hand, for given ε, if μ satisfies (3.11) and $v(x)$ satisfies

$$\int_0^{v(x)} \frac{1}{\sqrt{F(\mu) - F(\eta)}}\, d\eta = 2G(\mu) x,$$

on $0 < x \le \frac{1}{2}$ and $v(x) = v(1 - x)$ on $\frac{1}{2} < x < 1$, then it is easily seen that $v(0) = v(1) = 0, v(\frac{1}{2}) = \mu$, and

$$v'' + 2G^2(\mu)/(1 - v) = 0.$$

Repeating an argument similar to that leading to (3.8), we have

$$2G^2(\mu) = \varepsilon \|v\|^q,$$

which implies that $v(x)$ is a solution of (S). Hence the number of solutions of (3.11) is the same as that of (S).

To help determine the cardinality of the set of solutions of (3.11), we present some lemmas.

LEMMA 3.1. $\sqrt{2}G(\mu) > G(\lambda)$ *for* μ *in* $(0,1)$.

Proof. Since

$$G(\lambda) = \int_0^\lambda \frac{1}{\sqrt{F(\lambda) - F(\tau)}}\, d\tau \quad \text{and} \quad F(\lambda) = \log\frac{1}{1-\lambda} = \log\frac{1}{(1-\mu)^2} = 2F(\mu),$$

by using the change of variable $(1-\eta)^2 = 1 - \tau$, we obtain

$$G(\lambda) = \sqrt{2}\int_0^\mu \frac{1-\eta}{\sqrt{F(\mu) - F(\eta)}}\, d\eta,$$

and so

$$\sqrt{2}G(\mu) - G(\lambda) = \sqrt{2}\int_0^\mu \frac{\eta}{\sqrt{F(\mu) - F(\eta)}}\, d\eta > 0.$$

Since $G(\mu) > 0$ on $(0,1)$, it follows that $K(\mu) > 0$ on $(0,1)$. To investigate the behavior of $K(\mu)$ near $\mu = 0$ and $\mu = 1$, we use a transformation $\theta = \theta(\mu) = [\log(1/(1-\mu))]^{1/2}$. Since $\theta(\mu)$ is strictly increasing on $(0,1)$, we then have that for $\theta \in (0,\infty)$,

$$G(\mu) = 2e^{-\theta^2}\int_0^\theta e^{\sigma^2} d\sigma \tag{3.12}$$

and

$$G(\lambda) = 2e^{-2\theta^2}\int_0^{\sqrt{2}\theta} e^{\sigma^2} d\sigma. \tag{3.13}$$

By this change of variable, we find that

$$K(\mu) = H(\theta) = 2^{3+(q/2)}\left(e^{-\theta^2}\int_0^\theta e^{\sigma^2} d\sigma\right)^{2+q}\left(\sqrt{2}e^{-\theta^2}\int_0^\theta e^{\sigma^2} d\sigma - e^{-2\theta^2}\int_0^{\sqrt{2}\theta} e^{\sigma^2} d\sigma\right)^{-q}.$$

Then from L'Hôpital's rule

$$K(1) = \lim_{\theta\to\infty} H(\theta) = 0 \quad \text{for } q > 0 \tag{3.14}$$

and

$$K(0) = \lim_{\theta\to 0^+} H(\theta) = \begin{cases} 0, & \text{if } 0 < q < 1, \\ 12, & \text{if } q = 1, \\ +\infty & \text{if } q > 1. \end{cases} \tag{3.15}$$

LEMMA 3.2. *There is a* $\mu_0 \in (0,1)$ *such that* $1/\sqrt{F(\mu)} > G(\mu)$ *for* $0 < \mu < \mu_0$ *and* $1/\sqrt{F(\mu)} < G(\mu)$ *for* $\mu_0 < \mu < 1$.

Proof. $G(\mu)$ can be written in the form

$$G(\mu) = \int_0^\mu \frac{f(\eta)}{f(\eta)\sqrt{F(\mu) - F(\eta)}}\, d\eta,$$

where $f(\eta) = F'(\eta) = 1/1 - \eta$.

Upon integration by parts, we obtain

$$G(\mu) = 2\sqrt{F(\mu)} - 2\int_0^\mu \sqrt{F(\mu) - F(\eta)}\, d\eta. \tag{3.16}$$

Thus on $(0,1)$, $G'(\mu)$ exists and equals

$$\begin{aligned} G'(\mu) &= \frac{f(\mu)}{\sqrt{F(\mu)}} - f(\mu)\int_0^\mu \frac{1}{\sqrt{F(\mu) - F(\eta)}}\, d\eta \\ &= f(\mu)\left(\frac{1}{\sqrt{F(\mu)}} - G(\mu)\right). \end{aligned} \tag{3.17}$$

By (3.12), it is easy to check that $G(0) = G(1) = 0$. Since $G(\mu) > 0$ on $(0,1)$, there is at least a μ_0 with $G'(\mu_0) = 0$. If we can show that on $(0,1)$, $G'(\mu)$ has at most one zero, then it follows that $G'(\mu)$ is positive on $(0,\mu_0)$ and negative on $(\mu_0,1)$, and consequently, making use of (3.17) yields the desired result. To this end, we differentiate $G(\mu)$ twice to get

$$G''(\mu) = f'(\mu)\left(\frac{1}{\sqrt{F(\mu)}} - G(\mu)\right) + f(\mu)\left(-\frac{1}{2(F(\mu))^{3/2}} - \frac{1}{\sqrt{F(\mu)}} + G(\mu)\right).$$

At any point μ_0 where $G'(\mu) = 0$, we can see that $G''(\mu_0) = -f(\mu_0)/(2(F(\mu_0))^{3/2}) < 0$, which means $G(\mu)$ can only attain its maximum on $(0,1)$. Thus the proof is complete. □

Remark 3.1. A numerical computation gives $\mu_0 = 0.574$.

In order to analyze the monotonicity of $K(\mu)$, we turn our attention to $K'(\mu)$. By (3.17) and

$$\frac{d}{d\mu}G(\lambda(\mu)) = 2(1-\mu)f(\lambda)\left(\frac{1}{\sqrt{F(\lambda)}} - G(\lambda)\right) = 2f(\mu)\left(\frac{1}{\sqrt{2F(\mu)}} - G(\lambda)\right), \tag{3.18}$$

a straightforward calculation gives

$$\begin{aligned} K'(\mu) &= 2^{1+(q/2)}q(G(\mu))^{1+q}f(\mu) \\ &\quad\times\left\{\left(1+\frac{2}{q}\right)\left(\frac{1}{\sqrt{F(\mu)}} - G(\mu)\right)\left(\sqrt{2}G(\mu) - G(\lambda)\right)\right. \\ &\qquad\left. - G(\mu)\left(2G(\lambda) - \sqrt{2}G(\mu)\right)\right\} \\ &\quad\times(\sqrt{2}G(\mu) - G(\lambda))^{-(1+q)}. \end{aligned} \tag{3.19}$$

Define

$$J(\mu) = \left(1+\frac{2}{q}\right)\left(\frac{1}{\sqrt{F(\mu)}} - G(\mu)\right)(\sqrt{2}G(\mu) - G(\lambda)) - G(\mu)(2G(\lambda) - \sqrt{2}G(\mu)); \tag{3.20}$$

then it turns out that the sign of $K'(\mu)$ is the same as that of $J(\mu)$. However, the sign analysis of J is complicated. We can only obtain some incomplete results; others rely on numerical evidence.

Suppose that there is a $\mu_1 \in (0,1)$ such that $J(\mu_1) = 0$. Then a routine computation shows that at μ_1

$$J'(\mu) = f(\mu)\left\{ -\frac{1}{2}\left(1+\frac{2}{q}\right)(F(\mu))^{-3/2}(\sqrt{2}G(\mu) - G(\lambda)) + \left(\frac{2-q}{q}\right)G(\lambda)\left(\frac{1}{\sqrt{F(\mu)}} - G(\mu)\right)\right\}. \tag{3.21}$$

As a consequence, when $q = 2$, by Lemma 3.1 we see that $J'(\mu_1) < 0$, which implies that $K(\mu)$ should always attain its maximum on $(0,1)$. This is impossible since $\lim_{\mu\to 0^+} K(\mu) = +\infty$. Hence $J(\mu) < 0$ and $K(\mu)$ is decreasing for $0 < \mu < 1$ if $q = 2$.

Then, in view of Lemma 3.2, for $\mu_0 < \mu < 1$ and $q < 2$ or $0 < \mu < \mu_0$ and $q > 2$, we find

$$J(\mu) < 2\left(\frac{1}{\sqrt{F(\mu)}} - G(\mu)\right)(\sqrt{2}G(\mu) - G(\lambda)) - G(\mu)(2G(\lambda) - \sqrt{2}G(\mu)) < 0,$$

which shows that $K(\mu)$ is decreasing on $(\mu_0, 1)$ for $q < 2$ or on $(0, \mu_0)$ if $q > 2$.

We now write $J(\mu)$ in (3.20) as

$$J(\mu) = \frac{2}{q}\left(\frac{1}{\sqrt{F(\mu)}} - G(\mu)\right)\left(\sqrt{2}G(\mu) - G(\lambda)\right) + \frac{1}{\sqrt{F(\mu)}}(\sqrt{2}G(\mu) - G(\lambda) - \sqrt{F(\mu)}G(\mu)G(\lambda)), \tag{3.22}$$

and set

$$J_1(\mu) = \sqrt{2}G(\mu) - G(\lambda) - \sqrt{F(\mu)}G(\mu)G(\lambda). \tag{3.23}$$

LEMMA 3.3. *$J_1(\mu) < 0$ on $(0,1)$.*

Proof. Noting Lemmas 3.1 and 3.2 and the result for the case $q = 2$, we can see that $J_1(\mu) < 0$ on $(0, \mu_0]$. For $\mu_0 < \mu < 1$, we have

(3.24)

$$\begin{aligned} J_1'(\mu) &= f(\mu)\left(G(\lambda) - 2\sqrt{2}G(\mu) - \frac{1}{2\sqrt{F(\mu)}}G(\mu)G(\lambda) + 3\sqrt{F(\mu)}G(\mu)G(\lambda)\right) \\ &= f(\mu)(-2J_1(\mu) + G(\lambda)J_2(\mu)), \end{aligned}$$

where

$$J_2(\mu) = \left(\sqrt{F(\mu)} - \frac{1}{2\sqrt{F(\mu)}}\right)G(\mu) - 1, \tag{3.25}$$

with

$$J_2'(\mu) = f(\mu)\left(-J_2(\mu) + \frac{1}{2\sqrt{F(\mu)}}\left(G(\mu) - \frac{1}{\sqrt{F(\mu)}}\right) + \frac{G(\mu)}{4(F(\mu))^{3/2}}\right). \tag{3.26}$$

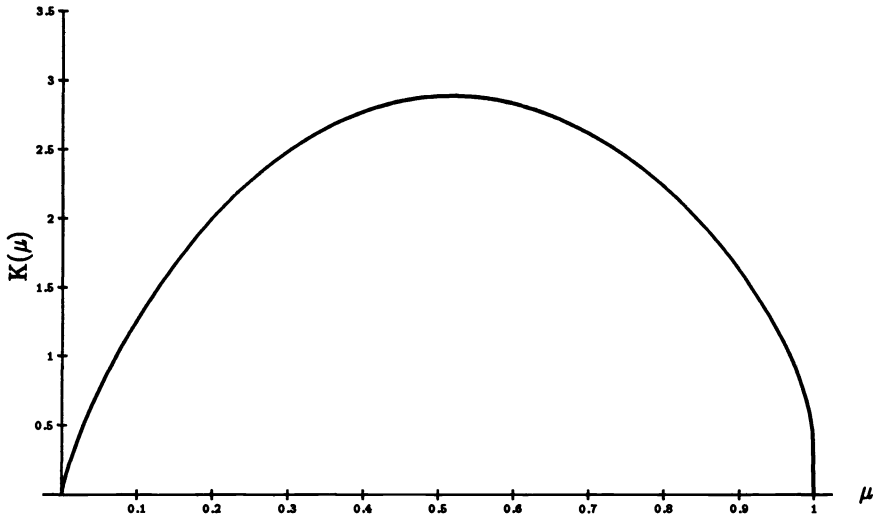


FIG. 1. $K(\mu)$ *for* $q = 0.2$.

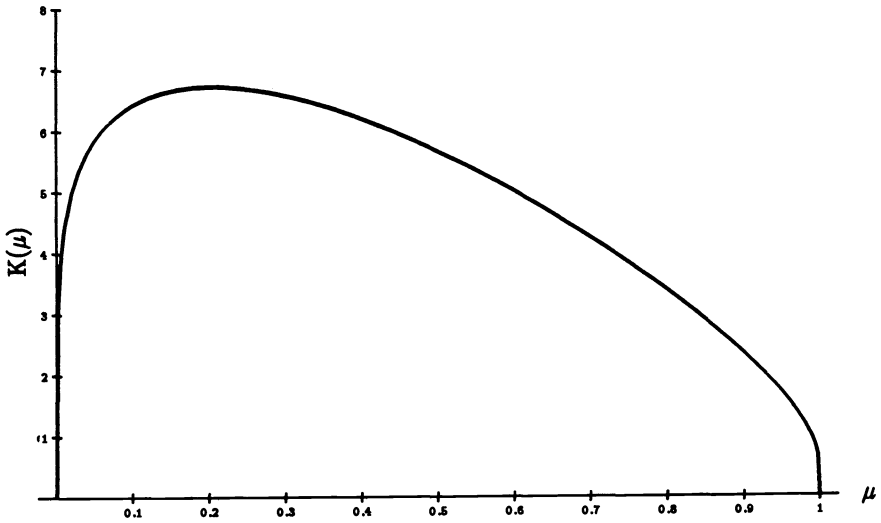


FIG. 2. $K(\mu)$ *for* $q = 0.8$.

If $J_1(\mu) \geq 0$ somewhere in $(\mu_0, 1)$, then there is at least one point $\mu_2(\mu_0 < \mu_2 < 1)$ such that $J_1(\mu_2) = 0$ and $J_1'(\mu_2) \geq 0$. It follows from (3.24) that $J_2(\mu_2) \geq 0$. Then by (3.26) and recalling Lemma 3.2, we can see that $J_2(\mu) > 0$ for all $\mu \in (\mu_2, 1)$, which contradicts the fact that $J_2 < 0$ near $\mu = 1$, since by L'Hôpital's rule $\lim_{\mu \to 1^-} J_2(\mu) = 0$ and $\lim_{\mu \to 1^-} J_2'(\mu) = +\infty$.

By means of (3.22) and all the lemmas, together with the fact that $J(\mu) < 0$ on $(0, \mu_0)$ if $q > 2$, we have the following theorem.

THEOREM 3.4. *If* $q \geq 2$, *there is a unique solution of* (S) *for each given* $\varepsilon > 0$.

Although we cannot provide a rigorous analysis in case $0 < q < 2$, the asymptotic formulas (3.15)–(3.16) and the numerical results of Figs. 1–3, allow us to assert the

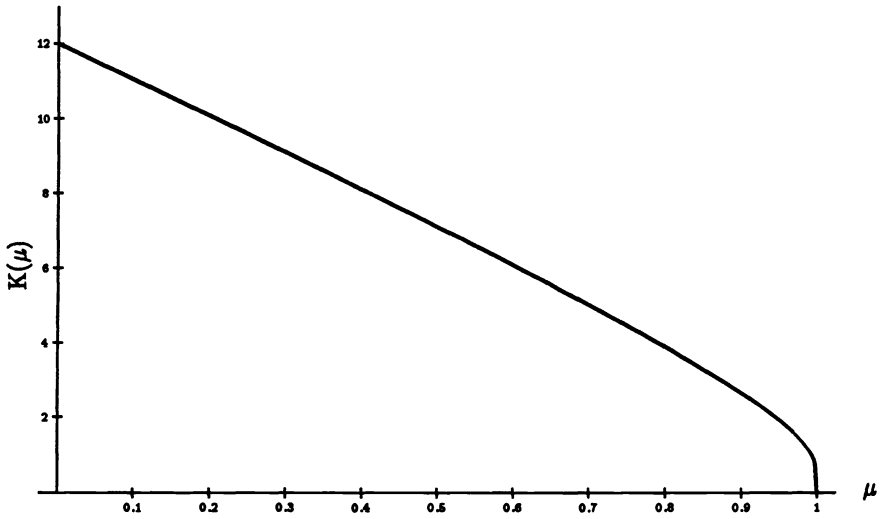


FIG. 3. $K(\mu)$ *for* $q = 1.0$.

following:

(C1) If $0 < q < 1$, there is an $\varepsilon(q)$ such that
  (a) if $\varepsilon > \varepsilon(q)$, there are no positive stationary solutions;
  (b) if $\varepsilon = \varepsilon(q)$, there is a unique solution; and
  (c) if $\varepsilon < \varepsilon(q)$, there are exactly two solutions.

(C2) If $q = 1$, we have the following:
  (a) if $\varepsilon \geq 12$, there are no positive stationary solutions; and
  (b) if $\varepsilon < 12$, there is exactly one solution.

(C3) If $1 < q < 2$, for every $\varepsilon > 0$, there is exactly one solution.

Remark 3.2. If one can verify (C2), then the validity of (C3) follows, since by (3.22) $J(\mu)$ is a decreasing function of q for $0 < \mu < \mu_0$. On the other hand, for $0 < q < 1$, it is easy to check that $\lim_{\mu \to 0+} K'(\mu) = +\infty$; hence (C1) (c) holds at least for small $\varepsilon > 0$.

Remark 3.3. In view of (3.1), (S) has no nonclassical solution $v_s(x)$ in the sense that $v_s(x)$ solves the equation in (S) classically on $(0,1)\backslash\{x = \frac{1}{2}\}$, but $v_s(\frac{1}{2}) = 1$.

4. Stability and quenching. Throughout this section, C1, C2, and C3 will be assumed. In order to get statements about the stability of solutions of (S) or the quenching result for problem (D), we first establish a relationship between solutions of (D) and those of (S).

LEMMA 4.1. *Suppose that* $u(x,t)$ *is monotone in* t *and* $\lim_{t \to \infty} u(x,t) = v(x)(<1)$ *exists. Then* $v(x)$ *is a classical solution of* (S).

Proof. Let

$$G(x,y) = \begin{cases} x(1-y), & 0 \leq x \leq y \leq 1, \\ y(1-x), & 0 \leq y \leq x \leq 1. \end{cases}$$

G is the Green's function for $-(d^2/dy^2)$ with Dirichlet boundary conditions. Since $u(x,t)$ is uniformly bounded and monotone in t, for fixed $x \in [0,1]$,

$$\lim_{t \to \infty} \int_0^1 G(x,y)u(y,t)\,dy = \int_0^1 G(x,y)v(y)\,dy,$$

and then

$$\frac{\partial}{\partial t}\int_0^1 G(x,y)u(y,t_n)\,dy \to 0$$

for some sequence $t_n \to \infty$. On the other hand,

$$\begin{aligned}\frac{\partial}{\partial t}\int_0^1 G(x,y)u(y,t)\,dy &= \int_0^1 G(x,y)u_t(y,t)\,dy \\ &= -u(x,t) + \varepsilon\|u(\cdot,t)\|^q \int_0^1 G(x,y)/(1-u(y,t))\,dy \\ &\to -v(x) + \varepsilon\|v\|^q \int_0^1 G(x,y)/(1-v(y))\,dy\end{aligned}$$

as $t \to \infty$.

Thus

$$v(x) = \varepsilon\|v\|^q \int_0^1 G(x,y)/(1-v(y))\,dy.$$

Moreover, since $v(x) < 1$ on $(0,1)$, $v(x)$ and $v'(x)$ are absolutely continuous, and hence $v(x)$ is a classical solution of (S).

By virtue of the above lemma, we can obtain a complete stability-instability result for stationary solutions of (D). This time, we treat the solution of (S) as a function of x depending on the parameter ε and denote it $v(x,\varepsilon)$.

LEMMA 4.2. *In the case* $0 < q < 1$, *if* $\varepsilon < \varepsilon(q)$, *the two solutions of* (S) *are ordered.*

Proof. Let $\mu_-(\varepsilon) < \mu_+(\varepsilon)$ and denote the corresponding solutions of (S) by $v_-(x,\varepsilon)$ and $v_+(x,\varepsilon)$, respectively. Assuming the assertion is not true, it follows that there is an $\bar{x} \in (0, \frac{1}{2})$ such that $v_-(\bar{x},\varepsilon) > v_+(\bar{x},\varepsilon)$. We now choose a $\delta(\delta < \varepsilon)$ so small that $v_-(x,\delta) \le v_-(x,\varepsilon)$ and $v_-(x,\delta) \le v_+(x,\varepsilon)$. Let $u(x,t,\varepsilon)$ be a solution of (D) with $u_0(x,\varepsilon) = v_-(x,\delta)$. Via the comparison theorem, we then find that

$$u(x,t,\varepsilon) \le v_-(x,\varepsilon) \quad \text{and} \quad u(x,t,\varepsilon) \le v_+(x,\varepsilon).$$

Moreover,

$$\begin{aligned}u_0'' + \varepsilon\|u_0\|^q/(1-u_0) &= v_-''(x,\delta) + \varepsilon\|v_-(\cdot,\delta)\|^q/(1-v_-(x,\delta)) \\ &> v_-''(x,\delta) + \delta\|v_-(\cdot,\delta)\|^q/(1-v_-(x,\delta)) \\ &= 0.\end{aligned}$$

Hence, recalling Corollary 2.2, we have $u_t(x,t,\varepsilon) \ge 0$ in D_T. By Lemma 4.1, $\lim_{t\to\infty} u(x,t,\varepsilon)$ exists and should be equal to one of the two solutions of (S), which leads to a contradiction, since $v_+(\frac{1}{2},\varepsilon) > v_-(\frac{1}{2},\varepsilon)$.

THEOREM 4.3. *Let* $v(x,\varepsilon)(<1)$ *be a positive solution of* (S) *on some* ε *interval* $[a,b]$, *and let* $\mu(\varepsilon) = v(\frac{1}{2},\varepsilon)$. *Then, if* $u'(\varepsilon) > 0$ *on* $[a,b]$, *the solution is stable, whereas it is unstable if* $\mu'(\varepsilon) < 0$.

Proof. For the case $\mu'(\varepsilon) > 0$ we first show that $v(x,\varepsilon_1) < v(x,\varepsilon_2)$ on $(0,1)$ for $a \le \varepsilon_1 < \varepsilon_2 \le b$.

From $\mu'(\varepsilon) > 0$, it follows that $\mu(\varepsilon_1) < \mu(\varepsilon_2)$ if $a \le \varepsilon_1 < \varepsilon_2 \le b$. In view of the conclusions in (C1), (C2), and (C3) in the previous section, we restrict our attention

to the case $0 < q < 1$. It is clear that $\mu'(\varepsilon) > 0$ can only occur in $(0, \mu_0)$, where μ_0 is the critical number in Lemma 3.2, since $K'(\mu) < 0$ in $[\mu_0, 1)$ if $0 < q < 2$. Noting the fact that $G'(\mu) > 0$ in $(0, \mu_0)$, by (3.4) we see that $\varepsilon_1 \|v(\cdot, \varepsilon_1)\|^q < \varepsilon_1 \|v(\cdot, \varepsilon_2)\|^q$. Then from (3.5), we find that for $0 < x \leq \frac{1}{2}$,

$$\int_0^{v(x,\varepsilon_1)} \frac{1}{\sqrt{F(\mu(\varepsilon_1)) - F(\eta)}}\, d\eta < \int_0^{v(x,\varepsilon_2)} \frac{1}{\sqrt{F(\mu(\varepsilon_2)) - F(\eta)}}\, d\eta, \tag{4.1}$$

and consequently

$$\int_{v(x,\varepsilon_1)}^{v(x,\varepsilon_2)} \frac{1}{\sqrt{F(\mu(\varepsilon_2)) - F(\eta)}}\, d\eta > 0.$$

Since $v(x, \varepsilon) = v(1-x, \varepsilon)$ if $\frac{1}{2} < x < 1, v(x, \varepsilon_1) < v(x, \varepsilon_2)$ on $(0, 1)$ for $a \leq \varepsilon_1 < \varepsilon_2 \leq b$.

Let $u(x, t, \varepsilon_1)$ be a solution of (D) with $u_0(x, \varepsilon_1) = v(x, \varepsilon_2)$. Then on $(0, 1)$, we have

$$\begin{aligned} u_0'' + \varepsilon_1 \|u_0\|^q/(1 - u_0) &= v_{xx}(x, \varepsilon_2) + \varepsilon_1 \|v(\cdot, \varepsilon_2)\|^q/(1 - v(x, \varepsilon_2)) \\ &< v_{xx}(x, \varepsilon_2) + \varepsilon_2 \|v(\cdot, \varepsilon_2)\|^q/(1 - v(x, \varepsilon_2)) \\ &= 0. \end{aligned}$$

Hence $u_t \leq 0$. From the comparison principle and monotonicity of u, on $(0, 1)$ we also have

$$v(x, \varepsilon_1) \leq u(x, t, \varepsilon_1) \leq v(x, \varepsilon_2).$$

By Lemma 4.1, $w(x, \varepsilon_1) = \lim_{t \to \infty} u(x, t, \varepsilon_1)$ is a solution of (S) satisfying $v(x, \varepsilon_1) \leq w(x, \varepsilon_1) \leq v(x, \varepsilon_2)$, and in particular, $v(\frac{1}{2}, \varepsilon_1) \leq w(\frac{1}{2}, \varepsilon_1) \leq v(\frac{1}{2}, \varepsilon_2)$. But w is a solution of (S) with ε_1, and so $w(x, \varepsilon_1)$ is either $v(x, \varepsilon_1)$ or else the other solution, $v_+(x, \varepsilon_1)$. From the graph of $K(\mu) = \varepsilon$, it follows that $v_+(\frac{1}{2}, \varepsilon_1) > v(\frac{1}{2}, \varepsilon_2)$, which excludes the possibility that $w(x, \varepsilon_1) = v_+(x, \varepsilon_1)$. We thus show that $v(x, \varepsilon_1)$ is stable from above. With $\varepsilon_1 > \varepsilon_2$, in a similar manner, we can also prove that $v(x, \varepsilon_1)$ is stable from below.

If $\mu'(\varepsilon) < 0$, we know that $v(x, \varepsilon_2) < v(x, \varepsilon_1)$ in a subinterval $[x_0, x_1]$ contained in $(0, 1)$ for $a \leq \varepsilon_1 < \varepsilon_2 \leq b$. Let $u(x, t, \varepsilon_2)$ be a solution of (D) with $u_0(x, \varepsilon_2) = v(x, \varepsilon_1)$. Then on $(0, 1)$, we find that

$$\begin{aligned} u_0'' + \varepsilon_2 \|u_0\|^q/(1 - u_0) &= v_{xx}(x, \varepsilon_1) + \varepsilon_2 \|v(\cdot, \varepsilon_1)\|^q/(1 - v(x, \varepsilon_1)) \\ &> v_{xx}(x, \varepsilon_1) + \varepsilon_1 \|v(\cdot, \varepsilon_1)\|^q/(1 - v(x, \varepsilon_1)) \\ &= 0. \end{aligned}$$

Thus $u_t \geq 0$ on D_T. Consequently, $u(x, t, \varepsilon_2)$ is increasing in t, which indicates that $v(x, \varepsilon_2)$ is unstable. Similarly, it can be shown that with $\varepsilon_1 > \varepsilon_2, v(x, \varepsilon_2)$ is also unstable.

Making use of this theorem combined with the characterization of the stationary solutions in §3, we state the following.

(S1) For $0 < q < 1$, there are two branches of solutions of (S), one stable, the other unstable.

(S2) For $q \geq 1$, the unique solution is unstable.

Next we establish quenching and global existence results for problem (D).

THEOREM 4.4. *Assuming the conjecture at the end of* §3, *we have the following results.*

(I) *Let* $0 < q < 1$.

(a) *For* $\varepsilon < \varepsilon(q)$, *if* $0 < u_0(x) < v_+(x,\varepsilon)$ *on* $(0,1)$, *then* u *exists for all* $t > 0$ *and* $\lim_{t\to\infty} u(x,t) = v_-(x,\varepsilon)$: *if* $u_0 > v_+(x,\varepsilon)$ *on* $(0,1)$, *then* u *quenches in finite time.*

(b) *For* $\varepsilon = \varepsilon(q)$, *if* $0 < u_0 \le v(x,\varepsilon(q))$, *then* u *is global and* $\lim_{t\to\infty} u(x,t) = v(x,\varepsilon(q))$; *if* $u_0 > v(x,\varepsilon(q))$, *then* u *is not global.*

(c) *For* $\varepsilon > \varepsilon(q)$, *if* $u_0(x) > 0$ *on* $(0,1)$, *every solution quenches in finite time.*

(II) *Suppose* $q = 1$.

(a) *For* $\varepsilon < 12$, *if* $0 \le u_0(x) < v(x,\varepsilon)$, *then quenching does not occur and* $\lim_{t\to\infty} u(x,t) = 0$, *while quenching happens in finite time if* $u_0(x) > v(x,\varepsilon)$.

(b) *For* $\varepsilon \ge 12$, *every solution with positive datum on* $(0,1)$ *quenches in finite time.*

(III) *Assume* $q > 1$. *For any* $\varepsilon > 0$, *if* $0 \le u_0(x) < v(x,\varepsilon)$, *then* u *exists globally and* $\lim_{t\to\infty} u(x,t) = 0$, *whereas* u *quenches in finite time if* $u_0(x) > v(x,\varepsilon)$.

Proof. Without causing any confusion, we shall sometimes write the solution of (D) with an initial value u_0 as $u(x,t;u_0)$.

(I) (a) We choose $v_0 = v_-(x,\sigma)$ with $\sigma < \varepsilon$ and $w_0 = v_+(x,\delta)$ with $\delta > \varepsilon$ such that $v_0 \le u_0 \le w_0$ on $(0,1)$. Then, by comparison,

$$u(x,t;v_0) \le u(x,t;u_0) \le u(x,t;w_0).$$

On the other hand, by recalling Corollary 2.2, we can see that $u(x,t;v_0)$ is monotonically increasing while $u(x,t;w_0)$ is monotonically decreasing, since $v_0'' + \varepsilon\|v_0\|^q/(1-v_0) > 0$ and $w_0'' + \varepsilon\|w_0\|^q/(1-w_0) < 0$. Hence, both $\lim_{t\to\infty} u(x,t;v_0)$ and $\lim_{t\to\infty} u(x,t;w_0)$ exist and equal $v_-(x,\varepsilon)$; consequently, so does $\lim_{t\to\infty} u(x,t;u_0)$.

If $u_0(x) > v_+(x,\varepsilon)$, we can find a δ ($\delta < \varepsilon$) so close to ε that $u_0(x) \ge v_+(x,\delta)$. Then set $v_0(x) = v_+(x,\delta)$. We observe that $u_t(x,t;v_0) \ge 0$ on the existence interval wherever u exists. Thus, if u does not quench for all $t > 0$, as $t \to \infty, u(x,t)$ tends to $w(x)$, a stationary solution of (D) with $w(\frac{1}{2}) > v_+(\frac{1}{2},\varepsilon)$. However, in view of Remark 3.3, there are no more stationary solutions of (D) even in the weak sense. Hence u must quench in finite time.

(b) The proof for (b) is similar to that for (a) and hence is omitted.

(c) We choose a δ with $\delta < \varepsilon$ such that $v_0(x) = v_-(x,\delta) \le u_0(x)$. Then we have $u(x,t;u_0) \ge u(x,t;v_0)$. Since $u_t(x,t;v_0) \ge 0, u(x,t;v_0)$ can only exist in finite time, and the conclusion follows.

(II) (a) Since $0 \le u_0(x) < v(x,\varepsilon)$, there is a number $\delta > \varepsilon$ such that $u_0(x) \le v(x,\delta)$ and $\mu(\delta) < \mu(\varepsilon)$. Let $w_0(x) = v(x,\delta)$. Then the comparison theorem yields that $0 \le u(x,t;u_0) \le u(x,t;w_0)$. Because $u(x,t;w_0)$ is monotonically decreasing, as $t \to \infty, u(x,t;w_0)$ approaches a stationary solution of (D) other than $v(x,\varepsilon)$, which must be the null solution.

(b) By the same reason as for (c) of (I), the proof is left out.

(III) We can argue in a similar manner, hence the proof is omitted.

Remark. In [8], for (K) it was proved that there is a critical number ε_0 such that when $\varepsilon < \varepsilon_0 (= \varepsilon_0)$, with certain data, the solution of (K) converges to the smaller (unique) stationary solution, while quenching occurs if $\varepsilon > \varepsilon_0$. The same result holds for solutions of a plasma-type equation $u_t = (u^m)_{xx} + \varepsilon/(1-u)(0 < m < 1)$ [2] and for those of a quasilinear parabolic equation $u_t = u_{xx}/(1+u_x^2) + 1/(1-u)$ [4]. For our problem, when $0 < q < 1$, similar dynamical behavior of solutions can be observed, whereas there are notable differences if $q \ge 1$. This shows that within necessary limits,

the influence of a nonlocal term can be so minimized that results for purely local problems are preserved for nonlocal ones, although analyses are more complicated in general.

5. Further discussion. In this section we describe briefly how the results can be generalized to other cases. First we point out that for problem (D), if the L_1 norm is replaced by the L_p norm ($p > 1$), a similar assertion can still hold. For instance, if $0 < q < 1$, we may conclude that there are two numbers ε_0 and ε_1 ($\varepsilon_0 < \varepsilon_1$) such that when $\varepsilon \le \varepsilon_0$, certain solutions exist globally, while finite time quenching happens if $\varepsilon > \varepsilon_1$. Since by the Hölder inequality, $\|u(\cdot,t)\|_1^q \le \|u(\cdot,t)\|_p^q$, it follows that the solution with the L_p norm is a supersolution of that with the L_1 norm. Hence, if $\varepsilon > \varepsilon_1 = \varepsilon(q)$, every solution must quench. On the other hand, following an idea in [1], we choose $v(x) = 2x(1-x)$ (the reason for such a choice was outlined in the proof of Theorem 4 in [1]) and find that

$$\max_{0\le x\le 1} \frac{1}{1-v} = 2 \quad \text{and} \quad \|v\|_p^q = 2^q \left[\int_0^1 x^p(1-x)^p\right]^{q/p} = 2^q[B(p+1,p+1)]^{q/p}.$$

Let $\varepsilon_0 = 2^{1-q}[B(p+1,p+1)]^{-q/p}$. Then if $\varepsilon \le \varepsilon_0, v(x)$ is a supersolution. Thus $u(x,t)$ is bounded from above by $v(x)$ if $u_0(x) \le v(x)$.

Next we indicate that arguments used for problem (D) can also apply to another kind of problem,

$$\text{(D*)} \qquad \begin{aligned} u_t &= u_{xx} + \varepsilon/(1-\|u(\cdot,t)\|^q u), && 0 < x < 1, \quad t > 0,\\ u(0,t) &= u(1,t) = 0, && t > 0,\\ u(x,0) &= u_0(x), && 0 \le x \le 1. \end{aligned}$$

Because the corresponding stationary solutions of (D^*) satisfy the following boundary value problem,

$$\text{(S*)} \qquad \begin{aligned} &v'' + \varepsilon/(1-\|v\|^q v), && 0 < x < 1,\\ &v(0) = v(1) = 0, \end{aligned}$$

we introduce a transformation $w(x) = \|v\|^q v$. The mapping $v \longmapsto w$ is one-to-one, since if $w_1(x) = w_2(x)$, it follows that $\|w_1\| = \|w_2\|$, which implies that $\|v_1\|^{q+1} = \|v_2\|^{q+1}$, and consequently, $v_1(x) = v_2(x)$. By such a transformation, (S^*) is changed into

$$\text{(S**)} \qquad \begin{aligned} &w'' + \varepsilon\|w\|^{q/(q+1)}/(1-w), && 0 < x < 1,\\ &w(0) = w(1) = 0. \end{aligned}$$

Since (S**) is of the same form as (S), we are able to conduct a discussion similar to our earlier one. However, it is interesting to note that $q/(q+1)$, the power of $\|w\|$, is always less than one. Thus we merely have assertion (C1), and consequently, only Theorem 4.4 (I) is valid. This illustrates that in problem (D), as a multiplier of the parameter ε, the nonlocal term makes a more powerful impact on the prevention of quenching.

Acknowledgment. The author would like to thank the referee for a number of valuable comments and constructive suggestions.

REFERENCES

[1] K. DENG, *The quenching of solutions of semilinear parabolic equations*, J. Huazhong Univ. Sci. Tech., 7 (1985), pp. 1–6.

[2] ———, *Quenching for solutions of a plasma type equation*, Nonlinear Anal. TMA, 18 (1992), pp. 731–742.

[3] K. DENG, M. K. KWONG, AND H. A. LEVINE, *The influence of nonlocal nonlinearities on the long-time behavior of solutions of Burgers' equation*, Quart. Appl. Math., 50 (1992), pp. 173–200.

[4] M. FILA, B. KAWOHL, AND H. A. LEVINE, *Quenching for quasilinear equations*, Comm. Partial Differential Equations, 17 (1992), pp. 593–614.

[5] H. KAWARADA, *On solutions of initial boundary value problem for* $u_t = u_{xx} + 1/(1-u)$, Res. Inst. Math. Ser., 10 (1975), pp. 729–736.

[6] H. A. LEVINE, *The phenomenon of quenching: A survey*, in Proc. 6th Internat. Conf. on Trends in the Theory and Practice of Nonlinear Analysis, North Holland, NY, 1985.

[7] ———, *Advances in quenching*, in Proc. Internat. Conf. on Reaction-Diffusion Equations and Their Equilibrium States, Birkhäuser, Boston, 1992.

[8] ———, *Quenching, nonquenching, and beyond quenching for solutions of some parabolic equations*, Ann. Mat. Pura Appl., 4, 155 (1989), pp. 243–260.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 112–128, January 1995

006

EXISTENCE AND BIFURCATION OF VISCOUS PROFILES FOR ALL INTERMEDIATE MAGNETOHYDRODYNAMIC SHOCK WAVES*

H. FREISTÜHLER† AND P. SZMOLYAN‡

Abstract. A viscous profile for a magnetohydrodynamic shock wave is given by a heteroclinic orbit of a six-dimensional gradient-like system of ordinary differential equations. This system, and thus possibly the existence of a viscous profile, vary with an array δ of four positive dissipation coefficients. It is known that for each choice of δ, all "classical" and "degenerate intermediate" shocks as well as some "nondegenerate intermediate" shocks have viscous profiles, and that, vice versa, each given nondegenerate intermediate shock has no viscous profile for some range of δ. Complementing this picture, it is shown that (i) each nondegenerate intermediate shock does have a (family of) viscous profile(s) for a certain other range of δ, and (ii) such profiles, for all intermediate shocks sharing the same relative flux, are generated in a global heteroclinic bifurcation. Both (i) and (ii) are proved in a regime of δ in which the dissipative effects due to electrical resistivity and longitudinal viscosity dominate those associated with transverse viscosity and heat conduction: The constructive proof is based on a recently formulated method in geometric singular perturbation theory.

Key words. shock waves, magnetohydrodynamics, heteroclinic orbits, singular perturbations

AMS subject classifications. 34C37, 34D15, 35L65, 76W05

1. Results. Under standard physical assumptions, plane magnetohydrodynamic waves are mathematically governed by the differential equations (e.g., [3]), in space $x \in \mathbf{R}$ and time $t \in \mathbf{R}$,

$$\begin{aligned} \rho_t + (\rho v)_x &= 0, \\ (\rho v)_t + (\rho v^2 + p + \tfrac{1}{2} \mid \mathbf{b} \mid^2)_x &= \lambda v_{xx}, \\ (\rho \mathbf{w})_t + (\rho v \mathbf{w} - \mathbf{b})_x &= \mu \mathbf{w}_{xx}, \\ \mathbf{b}_t + (v\mathbf{b} - \mathbf{w})_x &= \nu \mathbf{b}_{xx}, \\ \mathcal{E}_t + (v(\mathcal{E} + p + \tfrac{1}{2} \mid \mathbf{b} \mid^2) - \mathbf{w} \cdot \mathbf{b})_x &= \lambda (v v_x)_x + \mu (\mathbf{w} \cdot \mathbf{w}_x)_x + \nu (\mathbf{b} \cdot \mathbf{b}_x)_x + \kappa \theta_{xx}. \end{aligned} \tag{1.1}$$

Here $\rho, \theta, p(\rho, \theta) > 0$ denote density, temperature, and pressure of the fluid, $v \in \mathbf{R}$ its longitudinal and $\mathbf{w} \in \mathbf{R}^2$ its transverse velocity, $\mathbf{b} \in \mathbf{R}^2$ the transverse magnetic field, and

$$\mathcal{E} = \rho(\epsilon + \tfrac{1}{2}(v^2 + \mid \mathbf{w} \mid^2)) + \tfrac{1}{2} \mid \mathbf{b} \mid^2$$

the total energy with $\epsilon = \epsilon(\rho, \theta)$ the internal energy per mass. The "longitudinal" fluid viscosity λ and the "transverse" fluid viscosity μ are positive combinations ($\lambda = \zeta + \frac{4}{3}\eta, \mu = \eta$) of the two numbers $(\eta, \zeta \geq 0)$ known as the two viscosity coefficients (cf. [20, §§15, 49, and 78]) of the fluid, ν describes its electrical resistivity, and κ its heat conductivity; the terms involving these coefficients are macroscopic descriptions of dissipative mechanisms. We assume that apart from these mechanisms, the fluid is an ideal gas, i.e., pressure and internal energy are given by

$$p = R\theta\rho \quad \text{and} \quad \epsilon = c_v \theta,$$

* Received by the editors April 14, 1993; accepted for publication August 28, 1993.
† Institut für Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, D-52062 Aachen, Germany.
‡ Institut für Angewandte und Numerische Mathematik, Technische Universität Wien, A-1040, Austria.

where R and c_v, in principle arbitrary positive numbers, are the gas constant and the specific heat at constant volume, respectively. In addition to the thermodynamic quantities p and ϵ already introduced so far, we will also make use of the entropy

$$S = -R \ln \rho + c_v \ln \theta. \tag{1.2}$$

We abbreviate $(\rho, \rho v, \rho \mathbf{w}, \mathbf{b}, \mathcal{E}) \in U \subset (0, \infty) \times \mathbf{R} \times \mathbf{R}^2 \times \mathbf{R}^2 \times (0, \infty)$ as u, and $(\lambda, \mu, \nu, \kappa) \in [0, \infty)^4$ as δ, and write (1.1) briefly in the form

$$u_t + (f(u))_x = (D(u, \delta) u_x)_x. \tag{1.3}$$

In (1.3), all details of (1.1) are subsumed under the two real analytic mappings $f : U \to \mathbf{R}^7$ and $D : U \times [0, \infty)^4 \to \mathbf{R}^{7 \times 7}$.

In the limiting case $\delta = 0$, (1.3) is the (nonstrictly) hyperbolic system of conservation laws describing ideal magnetohydrodynamics. This case admits certain solutions of the form

$$u(x, t) = \begin{cases} u^-, & x < st, \\ u^+, & x > st; \end{cases} \tag{1.4}$$

they are precisely those functions of the form (1.4) which satisfy the Rankine–Hugoniot conditions, i.e.,

$$f(u^-) - su^- = f(u^+) - su^+ = q \tag{1.5}$$

for some constant q. A solution (1.4) of (1.3) with $\delta = 0$ is called a shock if both

$$m(S(u^+) - S(u^-)) > 0 \tag{1.6}$$

and

$$s \text{ is not an eigenvalue of } f'(u^-) \text{ nor of } f'(u^+) \tag{1.7}$$

hold, where in (1.6)

$$m = \rho^-(v^- - s) = \rho^+(v^+ - s) \in \mathbf{R} \tag{1.8}$$

is the relative mass flux associated with the solution (1.4), identical with the first component of the complete relative flux (1.5). Condition (1.6) means that entropy increases in the history of particles crossing the shock, a natural requirement based on the second law of thermodynamics. Condition (1.7) means that the discontinuity is noncharacteristic on either side.

For a given shock (1.4) and a given $\delta \in [0, \infty)^4, |\delta| > 0$, a heteroclinic orbit given by a solution $\phi : \mathbf{R} \to U$ of the system

$$D(\phi(x), \delta)\phi'(x) = f(\phi(x)) - s\phi(x) - q, \quad q \text{ from (1.5)}, \tag{1.9}$$

which connects the fixed points $u^\pm$:

$$\phi(\pm\infty) = u^\pm \tag{1.10}$$

is called a (δ-) profile of the shock. The idea of this notion is that the solution

$$u^\delta(x, t) = \phi(x - st) \tag{1.11}$$

of (1.3) is a regularized counterpart of (1.4) in which the effect of dissipation $D(.,\delta)$ is visibly resolved. The main result of this paper is the following theorem.

THEOREM 1.1. *Every magnetohydrodynamic shock has a δ profile at least for a certain open range of $\delta = (\lambda, \mu, \nu, \kappa) \in \Delta \equiv (0, \infty)^4$.*

Note that there are a number of different types of magnetohydrodynamic shock waves, and that for many q, several shocks of different types coexist with the same relative flux q.

Let us call a shock nondegenerate intermediate if, besides (1.4)–(1.7), it satisfies

$$|\mathbf{b}^+|\mathbf{b}^- = -|\mathbf{b}^-|\mathbf{b}^+ \neq 0, \tag{1.12}$$

i.e., the magnetic fields on either side are strictly antiparallel. All other shocks are classified as slow, fast, or degenerate intermediate. We mostly consider the nondegenerate intermediate shocks, because as we detail below, previously known results imply that all other shocks have δ profiles for all $\delta \in \Delta$. Theorem 1.1 can thus be viewed as a corollary of the following Theorem 1.2. To introduce a concise notation for the situation of interest, we will say that $q \in \mathbf{R}^7$ satisfies condition $\mathcal{I}$ iff a nondegenerate intermediate shock with relative flux q exists. Note also that if condition $\mathcal{I}$ holds for a given q, then all intermediate shocks with relative flux q are nondegenerate.

THEOREM 1.2. *Consider a $q \in \mathbf{R}^7$ which satisfies condition $\mathcal{I}$. Then there exist a number $\omega_0 > 0$ and a smooth function $\gamma : (\omega_0, \infty) \to (0, \infty)$ such that for all $\delta = (\lambda, \mu, \nu, \kappa)$ with*

$$\nu, \lambda > 0, \quad \nu/\lambda > \omega_0, \quad 0 \leq \mu/\lambda, \kappa/\lambda < \gamma(\nu/\lambda) \tag{1.13}$$

all shocks with relative flux q have δ profiles. In particular and more precisely, if $u^-, u^+ \in U, s \in \mathbf{R}$ define an intermediate shock and the number p, defined as the number of positive eigenvalues of $f'(u^-) - sI$ plus the number of negative eigenvalues of $f'(u^+) - sI$ minus eight, is positive, then there exists a p-parameter family of such profiles; if $p = 0$, then a pair of profiles exist.

We also show that the profiles for intermediate shocks are generated in a global heteroclinic bifurcation of system (1.9) with bifurcation parameter δ. A profile is called coplanar if the $\mathbf{b}$ and $\mathbf{w}$ components of ϕ' both remain in a fixed one-dimensional linear subspace of $\mathbf{R}^2$.

THEOREM 1.3. *Consider a $q \in \mathbf{R}^7$ which satisfies condition $\mathcal{I}$. Then there exist (possibly only small) numbers $\gamma_1, \gamma_2 > 0$ and a smooth function $\omega : [0, \gamma_1)^2 \to (0, \infty)$ such that the following holds for all $\delta = (\lambda, \mu, \nu, \kappa)$ with*

$$\nu, \lambda > 0, \quad 0 \leq \mu/\lambda, \kappa/\lambda < \gamma_1, \quad |\nu/\lambda - \omega(0,0)| < \gamma_2. \tag{1.14}$$

If $\nu/\lambda < \omega(\mu/\lambda, \kappa/\lambda)$, then no coplanar δ profiles exist, for any intermediate shock with relative flux q.

For $\nu/\lambda = \omega(\mu/\lambda, \kappa/\lambda)$, of all intermediate shocks with relative flux q only the (unique) one with $p = 0$ has a coplanar δ profile.

If $\nu/\lambda > \omega(\mu/\lambda, \kappa/\lambda)$, then all intermediate shocks with relative flux q (and $p > 0$) have (coplanar) δ profiles.

To interpret these theorems, we recall that slow and fast shocks have been proven in [12], [4], [5], [16] to possess (coplanar) δ profiles for all values of $\delta \in \Delta$ and many values $\delta \in \partial\Delta$. In [4] and [5] this was done by means of the Conley Index theory. For the marginal class of degenerate intermediate shocks the same is known (see [16] and [8]). It seems to be here that the assertion of Theorem 1.1 is established for the first

time for the class of nondegenerate intermediate shocks. Concerning these, [12], [4], and [5] had shown that for any nondegenerate intermediate shock there always is a range in Δ, namely of δ with $\lambda >> \mu, \nu, \kappa$ such that the shock has no δ profile. That nondegenerate intermediate shocks may conversely have $\tilde{\delta}$ profiles for other values $\tilde{\delta}$ ($\in \partial\Delta$), had been recognized long ago; cf. [18], [19, pp. 174–179], and [1]. In recent years Brio and Wu [2], and Wu ([23] and references therein), motivated by questions on the physics of the earth's magnetosphere, presented numerical evidence for stable magnetohydrodynamic traveling waves corresponding to such profiles. Connection matrix theory, a tool closely related to the Conley Index, was used in [21] to prove, nonconstructively, the existence of δ profiles, with *some* $\delta \in \Delta$, for *some* nondegenerate intermediate shocks. In [21], the authors stated also some conjectures about more profiles, which indeed amount to a picture that is roughly similar to our results. In [7] and [8] it was shown that for *all* $\delta \in \Delta$ there exist nondegenerate intermediate shocks which have δ profiles.

Complementing these earlier findings, the investigations whose results we report in this paper have had a double motivation. On one hand, it seems that proofs for the existence of viscous profiles represent a basic first step in any attempt at a more complete theoretical understanding of intermediate magnetohydrodynamic shock waves in the presence of dissipation. Also, we view our results more generally as a contribution to the ongoing debate on admissibility criteria (cf. [14]) in conservation law theory: Theorem 1.1 implies that in the case of magnetohydrodynamics—an example of *nonstrict* hyperbolicity—the simple criterion (1.6) of entropy increase has a satisfactory, consistent dynamical interpretation in the dissipative framework. This, in turn, contrasts especially with results which indicate that the stability of nondegenerate intermediate shock waves undergoes an explosive loss as $\delta \searrow 0$ even along rays $\delta/|\delta| = \text{const}$ (cf. [9], [10], and references therein). On the other hand, the o.d.e. system ((2.1) below) describing magnetohydrodynamic traveling waves is appealing from the dynamical systems point of view. The present paper is based on the idea of applying geometric singular perturbation theory (cf. [6], [22]) to this system.

2. Preliminaries and outline. We now turn to basic details of the problem and our approach. To determine the concrete form of (1.9), we rewrite (1.1), restricting attention to solutions which depend solely on x. The result is

$$\begin{aligned}
\rho v &= m,\\
\lambda \dot{v} &= \rho v^2 + p + \tfrac{1}{2} \mid \mathbf{b} \mid^2 - j,\\
\mu \dot{\mathbf{w}} &= \rho v \mathbf{w} - \mathbf{b} - \tilde{\mathbf{c}},\\
\nu \dot{\mathbf{b}} &= v\mathbf{b} - \mathbf{w} - \mathbf{c},\\
\kappa \dot{\theta} &= v(\mathcal{E} + p + \tfrac{1}{2} \mid \mathbf{b} \mid^2) - \mathbf{w} \cdot \mathbf{b} - \lambda v \dot{v} - \mu \mathbf{w} \cdot \dot{\mathbf{w}} - \nu \mathbf{b} \cdot \dot{\mathbf{b}} - e,
\end{aligned}$$

where dots denote differentiation with respect to x. The parameters m (see (1.8)), $j, e \in \mathbf{R}$, and $\mathbf{c}, \tilde{\mathbf{c}} \in \mathbf{R}^2$ correspond to a relative flux $q = (m, j, \tilde{\mathbf{c}}, \mathbf{c}, e)$. Using isotropy and Galilean invariance, we will from now on assume without loss of generality that

$$m > 0 \quad \text{and} \quad \tilde{\mathbf{c}} = 0.$$

We set $d = 1/m$, $\tau = v/m$ and rescale x by $1/m$; $\mathbf{w}$, $\mathbf{b}$ by m; θ, p, $\mathbf{c}$, j by m^2, and e by m^3 to obtain the following six-dimensional system, which we henceforth will refer to as Σ^6:

$$
\begin{aligned}
&\nu\dot{\mathbf{b}} = -d\mathbf{w} + \tau\mathbf{b} - \mathbf{c},\\
&\lambda\dot{\tau} = \tau + \frac{R\theta}{\tau} + \frac{1}{2} \mid \mathbf{b} \mid^2 - j,\\
&\mu\dot{\mathbf{w}} = \mathbf{w} - d\mathbf{b},\\
&\kappa\dot{\theta} = c_v\theta - \frac{1}{2}(\mid \mathbf{w} \mid^2 - 2d\mathbf{w}\cdot\mathbf{b} + \tau \mid \mathbf{b} \mid^2) - \frac{\tau^2}{2} + j\tau + \mathbf{b}\cdot\mathbf{c} - e.
\end{aligned}
\tag{2.1}
$$

Since in deriving (2.1) we had assumed independence of t, Σ^6 is (1.9) in the case $s = 0$. Again due to Galilean invariance, the restriction $s = 0$ implies no loss of generality. In other words, Σ^6 describes all profiles for magnetohydrodynamic shocks.

Before we outline and then enter the central part of our analysis, we show how previously known results imply the assertion of Theorem 1.1 in the cases which are not covered by Theorem 1.2.

LEMMA 2.1. *Theorem* 1.1 *holds if restricted to shocks whose relative flux* q *violates* $\mathcal{I}$.

To see this, note first the following lemma.

LEMMA 2.2. (i) *All gas-dynamic shocks have* δ *profiles for all values of* $\delta \in \Delta$. (ii) *All slow or fast shocks have* δ *profiles for all values of* $\delta \in \Delta$.

Proof. Statement (i) is a trivial consequence of [13]; compare also similar remarks in [7] and [8]. A gas-dynamic shock is one with $\mathbf{b}^- = \mathbf{b}^+ = 0$. Slow and fast shocks, unless they are gas-dynamic, are characterized by (non-anti- !) parallel magnetic fields, i.e.,

$$
|\mathbf{b}^+|\mathbf{b}^- = |\mathbf{b}^-|\mathbf{b}^+ \neq 0. \tag{2.2}
$$

The existence of profiles for such shocks is known from [4] and [12] and (ii) follows. □

Next observe the following lemma.

LEMMA 2.3. (i) *For any magnetohydrodynamic shock, properties* (1.5) *and* (1.7) *imply that the relative magnetic flux* $\mathbf{c}$ *occurring in the normalized system* Σ^6 *is different from* 0 *unless the shock is gas-dynamic.* (ii) *If* $q \in \mathbf{R}^7$ *is such that* $\mathbf{c} \neq 0$ *and* $\mathcal{I}$ *does not hold, then any shock with relative flux* q *is slow or fast.*

Proof. If $\mathbf{c} = 0$, fixed points of (2.1) with $\mathbf{b} \neq 0$—if there are any—occur grouped in whole circles. According to (1.7), however, the left- and right-hand states u^-, u^+ of a shock are isolated zeros of f. Thus only gas-dynamic shocks are possible if $\mathbf{c} = 0$. On the other hand, if $\mathbf{c} \neq 0$, any states u^-, u^+ satisfying (1.5) have $|\mathbf{b}^-|, |\mathbf{b}^+| > 0$ and fulfil thus either (2.2) or (1.12); violation of $\mathcal{I}$ precludes the latter. □

Remark. (i) Gas-dynamic shocks may be slow or fast or (degenerate!) intermediate.

(ii) Switch-on and switch-off discontinuities, characterized by (1.5), (1.6), and *either* $\mathbf{b}^- = 0$ *or* $\mathbf{b}^+ = 0$, and thus also having magnetic flux $\mathbf{c} = 0$, are no shocks in the sense of this paper: They violate our noncharacteristicity assumption (1.7). However, these discontinuities are already known [16] to have δ profiles for all $\delta \in \Delta$.

Lemmas 2.2 and 2.3 prove Lemma 2.1. Obviously, Lemma 2.1 and Theorem 1.2 together imply Theorem 1.1. Thus, we make the standing assumption that q satisfy $\mathcal{I}$ and prove the conclusions of Theorems 1.2 and 1.3. The whole rest of the paper is devoted to this purpose.

To approach it, we treat Σ^6 as a singularly perturbed problem: In the $\mu, \kappa << \nu, \lambda$ limit, the essential dynamics are captured by the three-dimensional reduced system

Σ^3 in the variables $\mathbf{b}$ and τ that one obtains from Σ^6 by setting $\mu = \kappa = 0$. We use methods from geometric singular perturbation theory [6] to prove the existence of a three-dimensional invariant manifold $\mathcal{C}$ for Σ^6 with the same dynamics as Σ^3. This reduces the whole problem to Σ^3. In our analysis of Σ^3, one key ingredient is the use of a Lyapunov function. While in the analysis of Σ^6 in [12] and [5] a mathematical potential P different from the physical entropy S has been used, it interestingly turns out that for Σ^3, S itself is a Lyapunov function. The other main ingredient in our analysis of Σ^3 consists in a further restriction of attention to the dynamics in the invariant half-plane $\mathbf{Rc} \times (0, \infty)$. The corresponding system Σ^2 in two variables was already discussed in [18]. Our analysis completes the reasoning of [18] and puts it on a mathematically sound basis. Summarizing, we can describe our strategy as finding heteroclinic orbits for Σ^2 and/or for Σ^3, and lifting them to Σ^6.

The geometric singular perturbation approach to the problem of viscous profiles for magnetohydrodynamic shock waves is new. In [5] and [16] slightly related ideas were used to construct isolating neighborhoods in an analysis for the case of limiting values of δ where some of its coefficients vanish. However, the logic in that approach is completely different from ours. The authors of [5] and [16] concluded (for fast and slow shocks) from the existence of δ profiles for all $\delta \in \Delta$ the existence of $\tilde{\delta}$ profiles for marginal dissipation $\tilde{\delta} \in \partial\Delta$. By contrast, we establish (for intermediate shocks) first $\tilde{\delta}$ profiles for certain $\tilde{\delta} \in \partial\Delta$ and subsequently induce δ profiles for a certain range of $\delta \in \Delta$.

Many of the new arguments we present below are not restricted to ideal gases. It would be interesting to investigate the same question for more general equations of state.

In the following §3, we carry out the reduction of Σ^6 to Σ^3. In §4, we analyze the geometry of Σ^3. Combining the results of §§3 and 4, the main results are proved in the final §5.

3. Geometric singular perturbation theory. The dynamical systems approach to singular perturbation problems—in its modern form—goes back to [6], but has only recently become more popular. In [22] a method—based on this invariant manifold approach—is formulated to prove the existence of transversal heteroclinic orbits of singularly perturbed differential equations. In this section we briefly summarize the necessary results from [6] and [22] and then apply them to completely reduce our analysis of the six-dimensional system Σ^6 to that of the three-dimensional system Σ^3. Consider first a general singularly perturbed system of differential equations in the standard form

$$\begin{aligned} \dot{\xi} &= X(\xi, \eta), \\ \varepsilon\dot{\eta} &= Y(\xi, \eta), \end{aligned} \tag{3.1}$$

with $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, $\varepsilon_0 > 0$ small, and $(\xi, \eta) \in U \subset \mathbf{R}^n \times \mathbf{R}^l$ open. We assume that $X : U \to \mathbf{R}^n$ and $Y : U \to \mathbf{R}^l$ are smooth functions. By setting $\varepsilon = 0$ we obtain the reduced problem

$$\begin{aligned} \dot{\xi} &= X(\xi, \eta), \\ 0 &= Y(\xi, \eta). \end{aligned} \tag{3.2}$$

The basic idea is to obtain orbits of the singularly perturbed problem (3.1), for small values of ε, as smooth perturbations of orbits of the reduced problem (3.2). The following results are contained in [6, Thm. 9.1].

THEOREM 3.1. *In addition to the assumptions made above in this section, assume that*

(a) *the equation* $Y = 0$ *has a manifold* $\mathcal{C}_0$ *of solutions which is the graph of a smooth function* $\hat{\eta} : \hat{U} \subset \mathbf{R}^n \to \mathbf{R}^l$ *and*

(b) *there exist integers* l_s *and* l_u, *with* $l_s + l_u = l$, *such that the partial Jacobian* Y_η *has* l_s *eigenvalues with negative real part, and* l_u *eigenvalues with positive real part, for all points of* $\mathcal{C}_0$.

Then the reduced problem (3.2) *defines a flow on* $\mathcal{C}_0$, *and the following assertions hold in an appropriate neighborhood of* $\mathcal{C}_0 \cap K$, *where* $K \subset U$ *is any compact set satisfying* $K \cap \mathcal{C}_0 \neq \emptyset$.

There exists $\varepsilon_1 > 0$ *such that* $\mathcal{C}_0$ *can be extended to a smooth family of manifolds* $\mathcal{C}_\varepsilon$, $\varepsilon \in (-\varepsilon_1, \varepsilon_1)$. *The manifolds* $\mathcal{C}_\varepsilon$ *are invariant under the flow of the singularly perturbed problem* (3.1), *and the restriction of this flow to* $\mathcal{C}_\varepsilon$ *is a smooth perturbation of the reduced flow on* $\mathcal{C}_0$. *If* $l_s = 0$ [$l_u = 0$], *then* $\mathcal{C}_\varepsilon$ *is moreover positively* [*negatively*] *isolated for the flow* (3.1), *for all* $\varepsilon \in (0, \varepsilon_1)$, *i.e., any orbit whose* ω- [α-] *limit set lies in* $\mathcal{C}_\varepsilon$ *lies itself completely in* $\mathcal{C}_\varepsilon$.

The above theorem is basic for the following proposition.

PROPOSITION 3.2. *Under the assumptions of Theorem* 3.1 *all structurally stable properties of the reduced problem* (3.2) *persist for the restriction of the singularly perturbed problem* (3.1) *to the invariant manifold* $\mathcal{C}_\varepsilon$ *for small* ε. *In particular, (normally) hyperbolic (manifolds of) fixed points of the reduced problem persist identically; their associated stable and unstable manifolds and transversal intersections of these perturb smoothly. Hyperbolicity and transversality extend to the singularly perturbed problem* (3.1) *(without restriction to* $\mathcal{C}_\varepsilon$*) with dimensions of stable and unstable manifolds increased by* l_s *and* l_u, *respectively.*

This proposition summarizes what we use from the second author's paper [22]. For details we refer to [6] and [22], for background material to [15] and [17]. Theorem 1.1 was announced in [22]. Another similar application of geometric singular perturbation theory can be found in [11].

We turn to applying geometric singular perturbation theory to the problem posed in §1. To make system Σ^6 accessible to Theorem 3.1 and Proposition 3.2, we fix $\nu, \lambda, \tilde{\mu}, \tilde{\kappa} > 0$ with $\tilde{\mu} + \tilde{\kappa} = 1$ arbitrarily and consider Σ^6 with $\mu = \varepsilon\tilde{\mu}, \kappa = \varepsilon\tilde{\kappa}$, for small values of $\varepsilon > 0$. From this point of view, Σ^6 is of the form (3.1) with $n = l = 3$ and $\xi = (\mathbf{b}, \tau), \eta = (\mathbf{w}, \theta)$. By setting $\varepsilon = 0$ we obtain the reduced problem Σ^3, which corresponds to (3.2):

$$
\begin{aligned}
\nu\dot{\mathbf{b}} &= -d\mathbf{w} + \tau\mathbf{b} - \mathbf{c}, \\
\lambda\dot{\tau} &= \tau + \frac{R\theta}{\tau} + \frac{1}{2} \mid \mathbf{b} \mid^2 - j, \\
\mathbf{w} &= d\mathbf{b} \equiv \hat{\mathbf{w}}(\mathbf{b}, \tau), \\
\theta &= \frac{1}{c_v}\left(\frac{1}{2}(\tau - d^2)|\mathbf{b}|^2 - \mathbf{b}\cdot\mathbf{c} + \frac{\tau^2}{2} - j\tau + e\right) \equiv \bar{\theta}(\mathbf{b}, \tau).
\end{aligned}
\tag{3.3}
$$

The last two lines describe the domain of definition

$$
\mathcal{C}_0 \subset U^6 \equiv \{(\mathbf{b}, \tau, \mathbf{w}, \theta) \in \mathbf{R}^2 \times (0, \infty) \times \mathbf{R}^2 \times (0, \infty)\}
\tag{3.4}
$$

of Σ^3. With $\hat{\theta} \equiv \bar{\theta}|U^3$, $\mathcal{C}_0$ is the graph of $\hat{\eta} = (\hat{\mathbf{w}}, \hat{\theta})$ over the reduced physical state space

$$
U^3 \equiv \{(\mathbf{b}, \tau) \in \mathbf{R}^3 \ : \ \tau > 0, \ \bar{\theta}(\mathbf{b}, \tau) > 0\};
\tag{3.5}
$$

this is assumption (a) of Theorem 3.1. Assumption (b) is also satisfied, with $l_s = 0$ and $l_u = 3$, since the 3×3 matrix $Y_\eta(\xi, \hat{\eta}(\xi))$, at each $\xi \in U^3$, has the eigenvalues 1, 1, and c_v, which are all positive. We restate [4, Thm. 4.1] as the next lemma.

LEMMA 3.3. *For each choice of* $(\mathbf{c}, d, e, j)$, *there is a compact ball* $K \subset U$ *which contains all fixed points and all heteroclinic orbits of* Σ^6 *with arbitrary* $\delta \in (\Delta \cup \partial\Delta)$.

Applying Theorem 3.1 with this K, for each quadruple $(\nu, \lambda, \tilde{\mu}, \tilde{\kappa})$ as above and each ε with $|\varepsilon| < \varepsilon_1$, $\varepsilon_1 > 0$ appropriate, we obtain an invariant manifold for Σ^6 with ν, λ as given and $\mu = \varepsilon\tilde{\mu}, \kappa = \varepsilon\tilde{\kappa}$. Analogous statements hold for the marginal cases $(\tilde{\mu}, \tilde{\kappa}) = (0, 1)$, $(\tilde{\mu}, \tilde{\kappa}) = (1, 0)$. By iterated application of Theorem 3.1 near these cases, we see that ε_1 can be chosen as $\varepsilon_1 = \gamma(\nu, \lambda) > 0$ independently of $\tilde{\mu}, \tilde{\kappa}$. Observing finally that the phase diagram of Σ^6 depends only on the mutual ratios of the coefficients $\nu, \lambda, \mu, \kappa$, i.e., briefly speaking, on $\delta/|\delta|$, we formulate the implications of Theorem 3.1 and Proposition 3.2 as the next lemma.

LEMMA 3.4. *There is a smooth function* $\tilde{\gamma} : (0, \infty) \to (0, \infty)$ *such that whenever* $\delta = (\lambda, \mu, \nu, \kappa) \in \Delta$ *satisfies* $\mu/\lambda, \kappa/\lambda < \tilde{\gamma}(\nu/\lambda)$, *then* Σ^6 *possesses a three-dimensional invariant manifold* $\mathcal{C}$, *located near* $\mathcal{C}_0$ *and depending smoothly on* $\delta/|\delta|$, *with the following properties:*

(i) *All fixed points and heteroclinic orbits of* Σ^6 *lie in* $\mathcal{C}$.

(ii) *Hyperbolic fixed points of* Σ^3 *are hyperbolic fixed points of* Σ^6 *with dimensions of the corresponding unstable manifolds increased by* 3.

(iii) *The existence of transversal (manifolds of) heteroclinic orbits connecting hyperbolic fixed points of* Σ^3 *implies the existence of transversal (manifolds of) heteroclinic orbits of* Σ^6 *in* $\mathcal{C}$ *connecting the same fixed points.*

By virtue of Lemma 3.4, we can henceforth restrict attention to Σ^3.

4. Geometry of the reduced problem. In this section, we study specific properties of Σ^3 which do not depend on the actual values of $\lambda, \nu > 0$ nor, within the limits of our overall assumption $\mathcal{I}$, on those of the quantities $\mathbf{c}, d, e, j$ associated with the relative flux q. Inserting the last two lines of (3.3) into its first two lines, Σ^3 obtains the form

$$\nu\dot{\mathbf{b}} = (\tau - d^2)\mathbf{b} - \mathbf{c}, \tag{4.1}$$

$$\lambda\dot{\tau} = \frac{1}{2}|\mathbf{b}|^2 + \tau - j + \frac{1}{k\tau}\left(-\frac{\tau^2}{2} - \frac{d^2}{2}|\mathbf{b}|^2 - \mathbf{b}\cdot\mathbf{c} + e\right), \tag{4.2}$$

where $k = 1 + c_v/R$ and λ denotes the original λ divided by $1 + R/c_v$. Mathematically, these equations define a smooth dynamical system $\overline{\Sigma^3}$ on the half-space $\tau > 0$ (of which Σ^3 is the restriction to U^3). Considering this extension in the sequel, we will thus sometimes be dealing with points, orbits, etc., which indeed lie outside the physical range U^3 by assuming nonpositive temperature $\bar{\theta}$. Note, however, that U^3 itself is positively invariant under the flow of $\overline{\Sigma^3}$, as we will prove immediately. First we observe what follows.

LEMMA 4.1. *Systems* Σ^3, $\overline{\Sigma^3}$ *are gradient-like: The entropy*

$$\hat{S}(\mathbf{b}, \tau) \equiv R \ln \tau + c_v \ln \hat{\theta}(\mathbf{b}, \tau), \quad (\mathbf{b}, \tau) \in U^3,$$

and/or the function

$$\overline{S}(\mathbf{b}, \tau) \equiv c_v \tau^{R/c_v} \bar{\theta}(\mathbf{b}, \tau), \quad (\mathbf{b}, \tau) \in \mathbf{R}^2 \times (0, \infty),$$

are strictly increasing along all nonstationary orbits.

Proof. A direct computation shows that

$$(\nu\dot{\mathbf{b}}, \lambda\dot{\tau})^\top = \hat{\theta}(\mathbf{b}, \tau)\nabla\hat{S}(\mathbf{b}, \tau)$$

and/or

$$(\nu\dot{\mathbf{b}}, \lambda\dot{\tau})^\top = \tau^{c_v/R}\nabla\overline{S}(\mathbf{b}, \tau)$$

hold for all $(\mathbf{b}, \tau) \in U^3$ or $(\mathbf{b}, \tau) \in \mathbf{R}^2 \times (0, \infty)$, respectively. □

Remark. $\hat{S}$ is also a reduced version of the mathematical potential P used in [12], [4], [5] and Lemma 4.1 follows thus partially also from (i) the existence of P and (ii) the fact that the reduced problem of a gradient-like singularly perturbed problem is gradient-like with respect to (the restriction of) the same Lyapunov function. Observations (i) and (ii) were made in [12] and [5], respectively.

LEMMA 4.2. *The physical range U^3 is invariant under the forward flow of $\overline{\Sigma^3}$.*

Proof. Consider the forward orbit O^+, with respect to the flow of $\overline{\Sigma^3}$, of a point $u \in \hat{U}$. $\hat{S}$ has a finite value at u and increases along O^+. Suppose that, along O^+, τ approaches zero. Then

$$\theta = \frac{1}{c_v}\left(\frac{1}{2}(\tau - d^2)|\mathbf{b}|^2 - \mathbf{b}\cdot\mathbf{c} + \frac{\tau^2}{2} - j\tau + e\right)$$

must become arbitrarily large, which is impossible. Now suppose that θ approaches zero along O^+. Then τ must become arbitrarily large, which implies that θ becomes large—a contradiction. Thus, τ and θ remain positive and bounded away from zero along O^+. □

Fix $\mathbf{e} \in \mathbf{R}^2$ such that

$$|\mathbf{e}| = 1, \qquad \mathbf{c} = c\mathbf{e} \quad \text{with } c \geq 0, \tag{4.3}$$

and let $E = \mathbf{R}\mathbf{e} \times (0, \infty) \subset \mathbf{R}^3$. The following is obvious.

LEMMA 4.3. *System $\overline{\Sigma^3}$ is invariant under reflection across E. In particular, E is invariant under the flow of $\overline{\Sigma^3}$.*

The properties of $\overline{\Sigma^3}$ to be discussed in the rest of this section are indeed properties of the restriction of $\overline{\Sigma^3}$ to E. Upon introducing the variable $b = \mathbf{b}\cdot\mathbf{e}$, $\overline{\Sigma^2} := \overline{\Sigma^3}|E$ and $\Sigma^2 := \Sigma^3|E$ are governed by the equations

$$\nu\dot{b} = (\tau - d^2)b - c, \tag{4.4}$$

$$\lambda\dot{\tau} = \frac{b^2}{2} + \tau - j + \frac{1}{k\tau}\left(-\frac{\tau^2}{2} - \frac{d^2}{2}b^2 - bc + e\right). \tag{4.5}$$

It is from now on that we make use of our overall assumption $\mathcal{I}$. By part (i) of Lemma 2.3, we have $\mathbf{c} \neq 0$; with (4.3), this amounts to

$$c > 0. \tag{4.6}$$

Note that $\mathbf{c} \neq 0$ implies that all fixed points of $\overline{\Sigma^3}$ must lie in E.

In the next step of our analysis we establish pertinent properties of the nullclines of $\overline{\Sigma^2}$. These are the portions in the half-plane $\tau > 0$ of the hyperbola $H \subset \mathbf{R}^2$ given by

$$h(b, \tau) \equiv (\tau - d^2)b - c = 0, \tag{4.7}$$

and of the solution set $G \subset \mathbf{R}^2$ of the equation

$$g(b,\tau) \equiv b^2(k\tau - d^2) - 2bc + (2k-1)\tau^2 - 2kj\tau + 2e = 0. \tag{4.8}$$

In particular, fixed points of $\overline{\Sigma^2}$ are, of course, given by elements of $H \cap G \cap (\mathbf{R} \times (0,\infty))$. Let H^-, H^+ denote the lower left ($\tau < d^2, b < 0$) and upper right ($\tau > d^2, b > 0$) branch of H, respectively.

LEMMA 4.4. *H and G intersect transversally in precisely four points. With (henceforth assumed) appropriate numbering, these points $u_i = (b_i, \tau_i), i = 0,1,2,3$ satisfy $H^+ \cap G = \{u_0, u_1\}, H^- \cap G = \{u_2, u_3\}$ and, more precisely,*

$$\tau_0 > \tau_1 > d^2 > \tau_2 > \tau_3 > 0,$$
$$b_1 > b_0 > 0 > b_3 > b_2.$$

At least for $i = 1$ and $i = 3$, u_i lies in the physical range $U^2 \equiv \{(b,\tau) \in \mathbf{R}^2 : \tau > 0$ and $\overline{\theta}(b\mathbf{e}, \tau) > 0\}$.

Proof. $H \cap G$ has at most four elements: Multiplying (4.8) by b^2 and substituting $\tau b = d^2 b + c$ from (4.7) yields a fourth-order polynomial in b. By virtue of $\mathcal{I}$ and (1.7), G intersects H^- and H^+ transversally in at least one point each. Since, however, $g \to \infty$ at each of the four infinities of H, $G \cap H^{\pm}$ consist then indeed both of two transversal intersection points. We number the altogether four points such that $b_1 > b_0 > 0 > b_3 > b_2$; this implies $\tau_0 > \tau_1 > d^2 > \tau_2 > \tau_3$.

It remains to show that u_1 and u_3 lie in U^2. Since, due to $\mathcal{I}$, $G \cap H^{\pm} \cap U^2 \neq \emptyset$, it suffices to note that $u_0 \in U^2$ would imply $u_1 \in U^2$, and $u_2 \in U^2$ would imply $u_3 \in U^2$. To see this, reduce $\overline{\Sigma^2}$ still further by setting $\nu = 0$:

$$\begin{aligned} 0 &= h(b,\tau), \\ \lambda\dot{\tau} &= g(b,\tau)/(2k\tau). \end{aligned} \tag{4.9}$$

The flow that these equations define on the portion of H lying in the half-plane $\tau > 0$ corresponds to the flow governed by the single equation

$$\lambda\dot{\tau} = g(c/(\tau - d^2), \tau)/(2k\tau) \tag{4.10}$$

on the two intervals $(0, d^2)$ and (d^2, ∞). In (d^2, ∞), this scalar flow has precisely two fixed points: τ_0 and τ_1. Since the right-hand side of (4.10) is negative for $\tau \in (\tau_1, \tau_0)$, this interval is a heteroclinic orbit from τ_0 to τ_1. Adding a second component $b = c/(\tau - d^2)$ lifts it to a heteroclinic orbit of (4.9). Since by an obvious extension of Lemma 4.2, the physical range $U^2 \cap H$ is positively invariant under the flow of (4.9), we conclude the desired implication $u_0 \in U^2 \Rightarrow u_1 \in U^2$. The argumentation for $u_2 \in U^2 \Rightarrow u_3 \in U^2$ is analogous once one observes additionally that τ_3 must be positive since $\tau_2 > 0 \geq \tau_3$ would again contradict the positive invariance of $U^2 \cap H$ under the flow of (4.9).

In the following we discuss the geometry of the nullclines H and G; cf. [18].

The hyperbola H has a horizontal asymptote at $\tau = d^2$, and the branches H^+ in $b > 0, \tau > d^2$ and H^- in $b < 0, \tau < d^2$. Thus, it remains to discuss G. For given τ the two possible solutions of (4.8) are given by

$$b = \frac{c \pm \sqrt{\pi(\tau)}}{k\tau - d^2}, \qquad \pi(\tau) = c^2 - ((2k-1)\tau^2 - 2kj\tau + 2e)(k\tau - d^2). \tag{4.11}$$

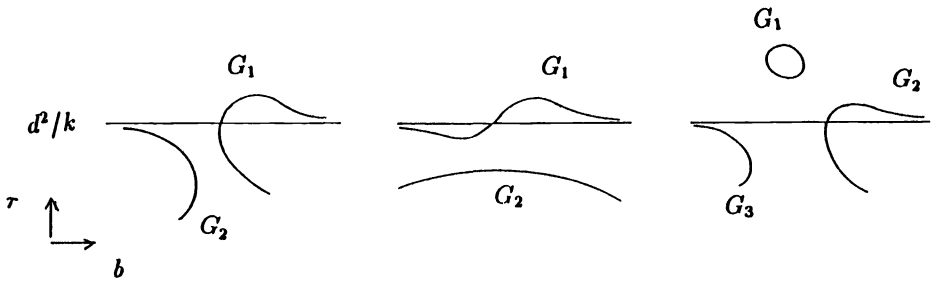


FIG. 1. *Nullclines* $g(b,\tau) = 0$ *in cases* (1), (2), *and* (3).

Real solutions exist for $\pi(\tau) \geq 0$. G has a horizontal asymptote at $\tau = d^2/k$, which G approaches from above as $b \to \infty$ and from below as $b \to -\infty$. Since $k > 1$, the horizontal asymptote of the hyperbola H lies above the horizontal asymptote of G. The geometry of G is determined by the number and the location of zeros τ_i^* of the cubic polynomial π. Since the leading coefficient of π is negative and $\pi(d^2/k) = c^2$ is positive, we have to distinguish three cases:

(1) there exists one zero $\tau_1^* > d^2/k$;
(2) there exist three zeros $\tau_1^* > d^2/k > \tau_2^* > \tau_3^*$;
(3) there exist three zeros $\tau_1^* > \tau_2^* > \tau_3^* > d^2/k$.

In cases (1) and (2), G has two connected components, G_1 and G_2; in case (3), G has three connected components G_1, G_2, and G_3, which are labeled in the order in which they appear as τ decreases (see Fig. 1). Note, that horizontal and vertical lines intersect G at most twice. Furthermore, G has exactly two, zero, and four vertical tangents in cases (1), (2), and (3), respectively. As the parameters c, d, e, and j vary, certain bifurcations are possible. Starting from case (2) the zeros τ_2^* and τ_3^* may coalesce and then disappear; from case (3) the zeros τ_1^* and τ_2^* or the zeros τ_2^* and τ_3^* may coalesce and then disappear; the resulting case is always case (1). Except at these bifurcations, which we do not discuss in the following but which can be treated similarly with unchanged results, G_1 and G_2 are separated by a vertical strip in case (1) and by a horizontal strip in case (2). In case (3), G_1 is separated from G_2 and G_3 by horizontal and vertical strips.

Now we discuss the relative position of H and G for the three cases. We know from Lemma 4.4 that H and G intersect transversally in four fixed points, $u_0, u_1 \in H^+$, $u_2, u_3 \in H^-$.

In case (1) there are two possibilities which give four fixed points.

(a) Both H^+ and H^- intersect G_1;
(b) H^+ intersects G_1, and H^- intersects G_2.

In case (2), remembering that the horizontal asymptote for H lies above the horizontal asymptote for G, we see that H^+ intersects G_1, and H^- intersects G_1 and G_2.

In case (3) there are four possibilities which give four fixed points.

(a) Both H^+ and H^- intersect G_1;
(b) H^+ intersects G_1 and H^- intersects G_3;
(c) both H^+ and H^- intersect G_2;
(d) H^+ intersects G_2 and H^- intersects G_3.

LEMMA 4.5. *The cases* (1.b), (3.b), *and* (3.d) *are not possible, i.e.,* H^- *does not intersect* G_2 *in case* (1), *and* H^- *does not intersect* G_3 *in case* (3).

Proof. We prove the lemma for case (1), the proof for case (3) is similar with G_2

replaced by G_3. G is the solution set of the equation $g(b,\tau)=0$. The critical points of g are the solutions of

$$g_b(b,\tau) = 2b(k\tau - d^2) - 2c = 0, \tag{4.12}$$

$$g_\tau(b,\tau) = kb^2 + 2(2k-1)\tau - 2kj = 0. \tag{4.13}$$

Equation (4.12) describes a hyperbola $\tilde{H}$, which is just H contracted by a factor k in the τ direction. In $b<0$, the parabola P defined by (4.13) intersects $\tilde{H}$ once in the critical point (b_*,τ_*), which is a saddlepoint. The topology of the level curves of g changes at the level value $g(b_*,\tau_*)$, in a way that corresponds to the transition between case (1) and case (2) upon variation of e. This implies that all of G_2 lies in the halfplane $b<b_*$. In particular, G_2 can intersect H^- at most in the quarterplane $Q=\{(b,\tau):\ b<b_*, \tau>\max\{\tau_*,0\}\}$. Since $g_\tau>0$ in Q and $g_b<0$ on $H^-\cap Q$, g is nowhere stationary along $H^-\cap Q$. Since both infinities of G_2 lie on the same side of H^-, this means that G_2 cannot intersect H^- at all. □

The above discussion and Lemma 4.5 imply that the three cases have the following common property.

LEMMA 4.6. *The set $G\cap([b_2,b_1]\times\mathbf{R})$ consists of two smooth graphs $G^\pm$ of functions $g^\pm:[b_2,b_1]\to\mathbf{R}$, distinguished by $g^-(b)<g^+(b), b\in(b_2,b_1)$. u_0 belongs to G^+, u_3 to G^-. u_1 and u_2 each lie on G^+ or G^- or both. (At least) in (b_2,b_1), both g^- and g^+ are smooth, and are stationary in at most one point.*

Finally we characterize the four fixed points.

LEMMA 4.7. (i) *u_0,u_1,u_2,u_3 are hyperbolic fixed points for the flow of $\overline{\Sigma^2}$; u_0 is an unstable node, u_3 is a stable node, u_1 and u_2 are saddles.*

(ii) *At u_1 and u_2 the stable and unstable manifolds are never tangent to vertical or horizontal lines, nor to G or H.*

(iii) *Interpreted, via suspension $(b,\tau)\mapsto(b\mathbf{e},\tau)$, as points in $\mathbf{R}^3$, u_0,u_1,u_2,u_3 are hyperbolic fixed points of $\overline{\Sigma^3}$. As such, the u_i have stable [unstable] manifolds of dimensions i $[3-i]$, $i=0,1,2,3$.*

Proof. For the reduced problem (4.9) u_0 and u_2 are repelling, u_1 and u_3 are attracting. Thus, for sufficiently small ν/λ assertion (i) follows by means of Proposition 3.2. We abbreviate (4.4) and (4.5) as

$$(\dot b,\dot\tau) = (h(b,\tau)/\nu, \tilde g(b,\tau)/\lambda)$$

and compute the derivative

$$A^{\nu,\lambda} = \begin{pmatrix} h_b/\nu & h_\tau/\nu \\ \tilde g_b/\lambda & \tilde g_\tau/\lambda \end{pmatrix}$$

of this vector field at any of the fixed points:

$$A^{\nu,\lambda}(b,\tau) = \begin{pmatrix} 1/\nu & 0 \\ 0 & \frac{1}{\lambda}(1-\frac{1}{k}) \end{pmatrix} \begin{pmatrix} \tau-d^2 & b \\ b & \frac{k}{k-1}(2-\frac{1}{k}+\frac{1}{\tau}(\frac{b^2}{2}-j)) \end{pmatrix}.$$

Taking into account that for symmetric matrices A, B with B positive definite, the number of negative and positive eigenvalues of the product BA does not depend on B, (i) follows for arbitrary $\nu,\lambda>0$. Assertion (ii) follows immediately upon multiplying $A^{\nu,\lambda}(b,\tau)$ by the vectors $(1,0)^\top$, $(0,1)^\top$, $(-h_\tau,h_b)^\top$, and $(-\tilde g_\tau,\tilde g_b)^\top$. To see (iii), we just note that in progressing from $\overline{\Sigma^2}$ to $\overline{\Sigma^3}$, an additional mode is added with eigenspace $E^\perp=\mathbf{e}^\perp\times\{0\}$ and eigenvalue $\tau-d^2$. □

Remarks. (i) Note that in connection with the signal speeds c_f, c_A, c_s of fast magnetoacoustic, Alfvén, and slow magnetoacoustic waves (cf. [3]; $c_f > c_A > c_s > 0$), assertion (iii) of Lemma 4.7 means that the physical states represented by the points u_0, u_1, u_2, u_3 correspond to flow with velocity $v > c_f$, $v \in (c_A, c_f)$, $v \in (c_s, c_A)$, and $v < c_s$, respectively. Correspondingly, the transition (u_0, u_1) is a fast shock, (u_2, u_3) a slow shock, and all (u_i, u_j) with $i \in \{0,1\}, j \in \{2,3\}$ are intermediate shocks.

(ii) The characterization of the fixed points of $\overline{\Sigma^2}$, $\overline{\Sigma^3}$ given in Lemmas 4.4 and 4.7 reflects well-known results [19] on the original full system Σ^6. The geometric statement of Lemma 4.6 puts us in a position to understand the dynamics of Σ^2 (see [18] and [19]) rigorously (see Lemma 5.1) and in all cases (see Fig. 1).

5. Existence and bifurcation of heteroclinic orbits. Gradient-like structure, positive invariance of physical state space, boundedness of the set of points on complete bounded orbits of positive temperature, the dimensions of stable and unstable manifolds, and specific geometric properties pose rather strong restrictions on the flow of Σ^3, which we will now use to prove the theorems stated in §1.

LEMMA 5.1. (i) *With a certain fixed ω, the two-dimensional system $\overline{\Sigma^2}$, depending on $\nu, \lambda > 0$, has heteroclinic orbits of the following types and no others:*

(a) $0 \to 1$, $0 \to 2$, $0 \to 3$, $1 \to 3$, $2 \to 3$ *for $\nu/\lambda > \omega$;*

(b) $0 \to 1$, $1 \to 2$, $2 \to 3$ *for $\nu/\lambda = \omega$;*

(c) $0 \to 1$, $2 \to 3$ *for $\nu/\lambda < \omega$.*

(ii) *At the bifurcation ratio $\nu/\lambda = \omega$, the unstable manifold of $\{u_1\} \times (0,\infty)^2$ and the stable manifold of $\{u_2\} \times (0,\infty)^2$, with respect to the extension of $\overline{\Sigma^2}$ by the equations $\dot{\nu} = 0$, $\dot{\lambda} = 0$, intersect transversally.*

(iii) *All orbits of types $0 \to 1$, $2 \to 3$, $0 \to 2$, $1 \to 3$, $1 \to 2$ are unique, while orbits of type $0 \to 3$ occur in a one-parameter family. In all cases there exist also orbits with α-limit u_0 [ω-limit u_3] which have no ω-limit [α-limit] in the physical range U^2.*

(iv) *The fixed points which lie in the physical range are ordered according to increasing values of the entropy S, i.e., $i < j$ implies $\hat{S}(u_i) < \hat{S}(u_j)$.*

(v) *The fixed points u_1, u_2, and u_3 always lie in the physical range U^2.*

Proof. (i) Consider the rectangle

$$R \equiv [b_2, b_1] \times (0, \bar{\tau}]$$

in the (b, τ) half-plane, with some $\bar{\tau} > \sup\{g^+(b); b_2 \le b \le b_1\}$ (see Fig. 2).

Independently of ν/λ, the flow of $\overline{\Sigma^2}$ leaves R through the boundary portion

$$(\partial R)^+ \equiv (\{b_2\} \times (\tau_2, \bar{\tau}]) \cup ([b_2, b_1] \times \{\bar{\tau}\}) \cup (\{b_1\} \times (\tau_1, \bar{\tau}]),$$

and enters R through

$$(\partial R)^- \equiv (\{b_2\} \times (0, \tau_2)) \cup (\{b_1\} \times (0, \tau_1)).$$

By part (ii) of Lemma 4.7, there are unique orbits Γ_i^u, Γ_i^s which approach the fixed points u_i, $i = 1,2$, from the interior of R, as their α or ω limit, respectively. We investigate possible positions of these orbits.

First, we show that assertions (a) and (c) hold in the extreme cases $\nu/\lambda >> 1$ and $\nu/\lambda << 1$, respectively. These cases can be treated by the geometrical singular perturbation theory from §3. For ν/λ sufficiently large there exists a one-dimensional attracting invariant manifold Γ_3 of $\overline{\Sigma^2}$, close to G^- and containing u_3. For the flow

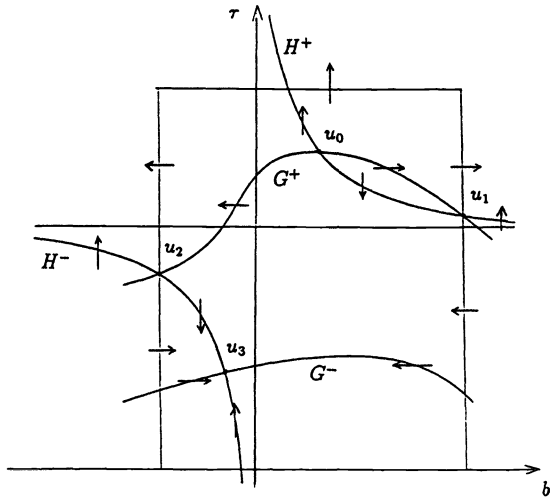


FIG. 2. *Phase portrait of* $\overline{\Sigma^2}$ *in* R.

restricted to Γ_3, u_3 is itself attracting. Let Γ_3^+ be the portion of Γ_3 in $b_3 < b < b_1$. Note that Γ_3^+ must lie in $\tau > 0$, since either the point u_1 or points on its almost vertical unstable manifold lie (arbitrarily) close to Γ_3 (for large enough ν/λ), which implies $\Gamma_3^+ \subset U^2$ by Lemmas 4.2 and 4.4. Hence, the almost vertical strongly unstable manifold of u_0 enters the domain of attraction of Γ_3^+ and an orbit $0 \to 3$ exists, as soon as $\nu/\lambda > \omega^*$ with ω^* sufficiently large—which we henceforth assume. (The idea underlying this argument consists in considering both the reduced and the "fast" problem into which $\overline{\Sigma^2}$ decouples for $\nu/\lambda \to \infty$; cf. [22].) This orbit $0 \to 3$ and parts of H and ∂R bound a region which Γ_1^u intersects but cannot leave. Thus, Γ_1^u has the ω limit u_3. The orbit Γ_1^u, the portion of H^- between u_2 and u_3, and $(\partial R)^+$ bound a region $R' \subset R$ which is negatively invariant. Since Γ_1^s and Γ_2^s point into R', they must both have an α limit u_0. Finally, as u_2 lies in the physical range U^2—this we anticipate from (v)—Γ_2^u is completely contained in $R \setminus (\mathrm{clos}(R'))$ and thus has ω-limit u_3. The four orbits Γ_1^s, Γ_1^u, Γ_2^s, Γ_2^u, being of type $0 \to 1$, $1 \to 3$, $0 \to 2$, $2 \to 3$, respectively, surround a region filled with orbits $0 \to 3$. We have shown that (a) holds if $\omega \geq \omega^*$.

From now on we quit the assumption $\nu/\lambda > \omega^*$ and consider instead cases where $\nu/\lambda < \omega_*$ with a possibly different number $\omega_* > 0$. If ω_* is sufficiently small, the orbits Γ_1^u, Γ_2^s are almost horizontal curves. This shows immediately that no heteroclinic orbit of type $i \to j$ exists, for any $(i,j) \in \{0,1\} \times \{2,3\}$. On the other hand, considering the portions of R above Γ_1^u and below Γ_2^s, respectively, we still see that Γ_1^s and Γ_2^u are heteroclinic orbits of type $0 \to 1$, $2 \to 3$, respectively. We have thus shown that (c) holds if $\omega \leq \omega_*$.

The phase portraits corresponding to cases (a) and (c) are both structurally stable (since they do not contain saddle–saddle connections). They are not equivalent (since, e.g., one of them contains an orbit of type $1 \to 3$, the other does not). Thus for at least a certain value $\omega \in [\omega_*, \omega^*]$, the phase portrait of $\overline{\Sigma^2}$ with $\nu/\lambda = \omega$ *does* contain an orbit Γ which is heteroclinic to saddles. Necessarily, $\Gamma = \Gamma_1^u = \Gamma_2^s$ and (b) holds with this ω. By part (ii) of Lemma 4.7, the intersections of Γ with sufficiently small

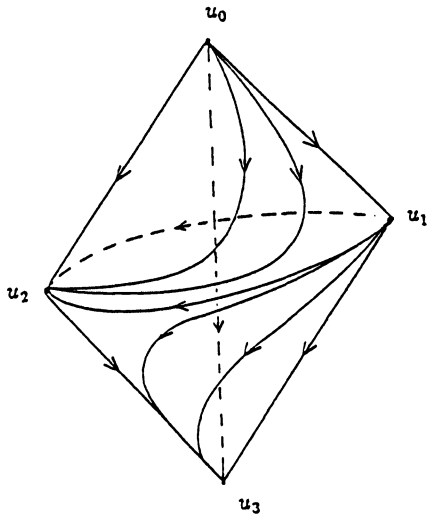


FIG. 3. *Connecting orbits of* Σ^3.

neighborhoods of u_1 and u_2 lie inside the region

$$N \equiv \{(b,\tau) \in R : g(b,\tau) < 0, h(b,\tau) < 0\}.$$

The vertical or horizontal orientation of the vector field along H and G, respectively, together with the piecewise monotonicity property of $g^{\pm}$ mentioned in Lemma 4.6 imply that Γ lies completely in N. At least for ν/λ in a neighborhood of ω, the orbits Γ_1^u, Γ_2^s intersect the $b = 0$ line in unique points $(0,\tau_1^u), (0,\tau_2^s)$ where τ_1^u and τ_2^s are smooth functions of ν/λ. Noting that $\Gamma \subset N$ implies $(\tau_1^u - \tau_2^s)'(\omega) < 0$, we see that the bifurcation takes place locally near ω as described in (a)–(c). Hence ω is unique and (a)–(c) hold globally.

(ii) Immediate from $(\tau_1^u - \tau_2^s)'(\omega) \neq 0$.

(iii) Follows directly from the facts that Γ_1^s connects to u_0, Γ_2^u to u_3, and the behavior of the vector field along H.

(iv) Follows from (i) and Lemma 4.1.

(v) From Lemma 4.4 we know already that $u_1, u_3 \in U^2$. The facts that $u_1 \in U^2$ and that a $1 \to 2$ orbit exists for $\nu/\lambda = \omega$ (case (b)) imply that also $u_2 \in U^2$, by virtue of Lemma 4.2. □

The following lemma characterizes the flow of Σ^3.

LEMMA 5.2. (i) *Assume that* Σ^3 *has four fixed points and that an orbit* $0 \to 3$ *exists in* E. *Then, unique orbits* $0 \to 1$, $2 \to 3$, *a pair of orbits* $1 \to 2$, *and one-parameter families of orbits* $0 \to 2$, $1 \to 3$ *exist for* Σ^3. *The union of these orbits and the fixed points is the boundary of a two-parameter family of orbits* $0 \to 3$. (ii) *Assume that* Σ^3 *has the three fixed points* u_1, u_2, *and* u_3 *and that an orbit* $1 \to 3$ *exists. Then, a unique orbit* $2 \to 3$, *a pair of orbits* $1 \to 2$, *and a one-parameter family of orbits* $1 \to 3$ *exist for* Σ^3.

The situation in the case of assertion (i) is shown in Fig. 3.

Proof. We consider a sufficiently small sphere Ω centered at u_3 and on this sphere a maximal connected reflectionally invariant set $\Omega_0 \neq \emptyset$ of points with α limit u_0.

All orbits passing through points in Ω_0 are heteroclinic orbits $0 \to 3$. We denote the union of these orbits by B.

B is an open, bounded subset of $\mathbf{R}^3$. The invariance of B implies that its boundary ∂B is invariant. Consider now the boundary $\partial\Omega_0$ of Ω_0 relative to Ω. Any point in $\partial\Omega_0$ has u_3 as its ω limit and must have u_1 or u_2 as its α limit . Since certain other points of Ω have no α limit inside the physical domain U^3, $\partial\Omega_0$ contains more than one point. Since the unstable manifold of u_2 lies in the invariant plane E, there is at most one orbit $2 \to 3$; hence, $\partial\Omega_0$ contains in particular a point with α limit u_1, i.e., an element of the unstable manifold W_1^u of u_1. Thus, $\partial\Omega_0$ contains a maximal smooth curve $\Gamma \subset \Omega \cap W_1^u$ of such points. Γ is either a closed curve or has endpoints. If Γ were a closed curve, then all points in B would have α limit u_1, an obvious contradiction. The endpoints of Γ have an α limit u_2 and coincide, lying on the thus existing and necessarily unique orbit $2 \to 3$. The nonempty boundary, with respect to W_1^u, of the set $\partial B \cap W_1^u$ consists of orbits $1 \to 2$; by the reflectional symmetry of Σ^3 (see Lemma 4.3) it is a pair of such orbits.

The remaining assertions of (i) follow by reversing the direction of the flow and the rôles of u_0, u_1, u_2, and u_3. Statement (ii) is proved in the same way. □

Remark. Note that the proof of Lemma 5.2 uses only (parts of) assertions (iii)–(v) of Lemma 5.1—and not the detailed knowledge statements (i) and (ii) provide about bifurcation. The following observation yields a closer connection between the statements of the two lemmas. For $\nu/\lambda = \omega$, the two-dimensional manifold W_1^u of u_1 intersects the two-dimensional stable manifold W_2^s of u_2 in the orbit $1 \to 2$, which lies in the invariant subspace E. Due to the reflectional symmetry of $\overline{\Sigma^3}$ with respect to E, the intersection of W_1^u and W_2^s is nontransversal. If W_1^u and W_2^s are in a sufficiently general position, e.g., if the intersections of W_1^u and W_2^s with the plane $b = 0$, for $\nu/\lambda = \omega$, have contact of some finite order, then a pair of symmetrically located transversal orbits $1 \to 2$ of $\overline{\Sigma^3}$ is generated as ν/λ passes through ω. Since the orbits $0 \to 1$, $2 \to 3$ are trivially transversal, the λ-Lemma [15] then implies the existence of the families of orbits $0 \to 2$, $1 \to 3$, $0 \to 3$ described in Lemma 5.2 for $\nu/\lambda - \omega$ small and positive. However, it seems difficult to verify the generic condition of sufficiently general position analytically.

All orbits established in Lemma 5.2—with the possible exception of orbits $1 \to 2$—are transversal, because of the dimensions of the corresponding stable and unstable manifolds. Since by Lemma 5.1 an orbit $0 \to 3$ (or $1 \to 3$ if u_0 is not physical) exists for $\nu/\lambda > \omega$, Lemma 5.2, Lemma 3.4, Proposition 3.2, and the analogue of Lemma 5.2 for the restriction of Σ^6 to the invariant manifold $\mathcal{C}$ imply the assertion of Theorem 1.2.

To prove Theorem 1.3 we consider systems Σ^4, $\overline{\Sigma^4}$ of four equations

$$\begin{aligned}
\nu\dot{b} &= -dw + \tau b - c,\\
\lambda\dot{\tau} &= \tau + \frac{R\theta}{\tau} + \frac{1}{2} \mid b \mid^2 - j,\\
\mu\dot{w} &= \quad w - db,\\
\kappa\dot{\theta} &= c_v\theta - \frac{1}{2}(\mid w \mid^2 - 2dw \cdot b + \tau \mid b \mid^2) - \frac{\tau^2}{2} + j\tau + bc - e,
\end{aligned}$$

where $(b, \tau, w) \in \mathbf{R} \times (0, \infty) \times \mathbf{R}$ and $\theta > 0$ or $\theta \in \mathbf{R}$, respectively. Representing the restriction of Σ^6 to $E \times E$, Σ^4 governs all coplanar profiles and is related to Σ^2 in precisely the same way as Σ^6 is to Σ^3. Thus, Proposition 3.2, the analogue of Lemma 3.4 for Σ^4, and Lemma 5.1, in particular its part (ii), imply that the bifurcation of

heteroclinic orbits of Σ^4 occurs as described in Theorem 1.3.

Acknowledgment. We would like to acknowledge the hospitality of the Institute for Mathematics and Its Applications, Minneapolis, where we carried out part of these investigations in 1989 and 1990.

REFERENCES

[1] J. E. ANDERSON, *Magnetohydrodynamic shock waves*, MIT Press, Cambridge, MA, 1963.
[2] M. BRIO AND C. C. WU, *An upwind differencing scheme for the equations of magnetohydrodynamics*, J. Comput. Phys., 75 (1988), pp. 400–422.
[3] H. CABANNES, *Theoretical magnetofluiddynamics*, Academic Press, New York, 1970.
[4] C. CONLEY AND J. SMOLLER, *On the structure of magnetohydrodynamic shock waves*, Comm. Pure Appl. Math., 27 (1975), pp. 367–375.
[5] ———, *On the structure of magnetohydrodynamic shock waves*, II, J. Math. Pures Appl., 54 (1975), pp. 429–444.
[6] N. FENICHEL, *Geometric singular pertubation theory*, J. Differential Equations, 31 (1979), pp. 53–98.
[7] H. FREISTÜHLER, *Linear degeneracy and shock waves*, Math. Z., 207 (1991), pp. 583–596.
[8] ———, *Some remarks on the structure of intermediate magnetohydrodynamic shocks*, J. Geophys. Res., 96 (1991), pp. 3825–3827.
[9] ———, *Non-uniformity of vanishing viscosity approximation*, Appl. Math. Lett., 6 (1993), pp. 35–41.
[10] H. FREISTÜHLER AND T.-P. LIU, *Nonlinear stability of overcompressive shock waves in a rotationally invariant system of viscous conservation laws*, Comm. Math. Phys., 153 (1993), pp. 147–158.
[11] I. GASSER AND P. SZMOLYAN, *A geometric singular perturbation analysis of detonation and deflagration waves*, SIAM J. Math. Anal., 24 (1993), pp. 968–986.
[12] P. GERMAIN, *Contribution à la théorie des ondes de choc en magnétodynamique des fluides*, O.N.E.R.A. Publ. No. 97 (1959).
[13] D. GILBARG, *The existence and limit behavior of the one-dimensional shock layer*, Amer. J. Math., 73 (1951), pp. 256–274.
[14] J. GLIMM, *The interaction of nonlinear hyperbolic waves*, Comm. Pure Appl. Math., 41 (1988), pp. 569–590.
[15] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[16] M. HESARAAKI, *The structure of shock waves in magnetohydrodynamics*, Amer. Math. Soc. Mem., 302 (1984).
[17] M. W. HIRSCH, C. C. PUGH, AND M. SHUB, *Invariant manifolds*, Springer Lecture Notes in Math., Vol. 583, Springer-Verlag, New York, Berlin, Heidelberg, 1979.
[18] A. KULIKOVSKII AND G. LIUBIMOV, *On the structure of an inclined magnetohydrodynamic shock wave*, Appl. Math. Mech., 25 (1961), pp. 171–179.
[19] A. KULIKOVSKII AND G. LIUBIMOV, *Magnetohydrodynamics*, Addison-Wesley, Reading, MA, 1965.
[20] L. D. LANDAU AND E. M. LIFSHITZ, *Fluid Mechanics*, Pergamon Press, Oxford, 1959.
[21] K. MISCHAIKOW AND H. HATTORI, *On the existence of intermediate magnetohydrodynamic shock waves*, J. Dynamics Diff. Eqs. 2 (1990), pp. 163–175.
[22] P. SZMOLYAN, *Transversal heteroclinic and homoclinic orbits in singular pertubation problems*, J. Differential Equations 92 (1991), pp. 252–281.
[23] C. C. WU, *Formation, structure, and stability of MHD intermediate shocks*, J. Geophys. Res. 95 (1990), pp. 8149–8175.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 129–146, January 1995

007

TIME-LIKE TRACE REGULARITY OF THE WAVE EQUATION WITH A NONSMOOTH PRINCIPAL PART*

GANG BAO† AND WILLIAM W. SYMES‡

Abstract. A trace regularity result is established for the multidimensional wave equation with nonsmooth variable coefficients in the principal part. It is shown that the time like trace of the solution can be as regular as the solution itself, provided that microlocal restrictions against the tangential oscillations of the coefficients.

Key words. trace regularity, microlocal Sobolev spaces, nonsmooth symbol classes, propagation of singularities, pseudodifferential cutoff

AMS subject classifications. 35L10, 35R25, 35S05

1. Introduction.

A simplified model that governs many physical processes, such as acoustic and seismic wave propagation, is the following wave equation:

$$\left[\frac{1}{c^2(x,t)}\frac{d^2}{dt^2} - \nabla\sigma(x,t)\cdot\nabla\right] u(x,t) = f(x,t),$$

where both coefficients $c(x,t)$ and $\sigma(x,t)$ are functions that may or may not be smooth. In this paper, we continue our study of trace regularity properties for the solution of the multidimensional wave equation with nonsmooth coefficients. The classical trace theorem in Sobolev spaces indicates that there will be a half-derivative loss when a distribution is being restricted to a codimension one hypersurface. On the other hand, the standard method of energy estimates yields that for a strictly hyperbolic partial differential equation with smooth coefficients, the restriction (or trace) map to a codimension one space-like hypersurface, mapping the solution to its trace, is from $H^s(\mathbf{R}^k)$ to $H^s(\mathbf{R}^{k-1})$ locally for any real s. It is our goal in this work to investigate the circumstances under which a time-like trace of the solution is as regular as the solution itself. Obviously, even with smooth coefficients, the answer would be negative without any assumptions, which is essentially due to the ill-posedness nature of time-like hyperbolic Cauchy problems, or the presence of grazing rays. Moreover, the situation will be more complex when nonsmooth coefficients are present, where only limited initial regularity can be propagated.

In many applications, e.g., control and inverse problems, the situation where the coefficients (i.e., the medium) are time independent is of particular interest. Thus we shall throughout assume that the coefficients are time free in the model, i.e., $c(x,t) = c(x)$ and $\sigma(x,t) = \sigma(x)$, although some of the analysis may be extended to the time-dependent case.

* Received by the editors February 16, 1993; accepted for publication (in revised form) August 31, 1993.

† Department of Mathematics, University of Florida, Gainesville, Florida 32611. The research of this author was partially supported by the National Science Foundation through a grant to the Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, Minnesota, and a grant from Honeywell, Inc.

‡ Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77251-1892. The research of this author was partially supported by National Science Foundation grant DMS 8905878, Office of Naval Research contract N00014-89-J-1115, Air Force Office of Scientific Research grant 90-0334, the Texas Geophysical Parallel Computation Project, and the Rice Inversion Project.

The following theorem is the main result of this paper. We show that with additional microlocal smoothness along the tangential directions in the coefficients, the time-like trace can be as regular as the solution itself. Let us denote $d^2(x) = 1/c^2(x)$ and $\Box_d = d^2(x)\partial_t^2 - \Delta$ and $\xi = (\xi', \xi_n)$ is the Fourier variable corresponding to $x = (x', x_n)$. Throughout, the function $c(x)$ is always assumed to be positive as it stands for wave speed in applications.

THEOREM 1.1. *Suppose that $s > 3 + n/2$ and that u solves the problem*

$$[\Box_d - \nabla\sigma(x) \cdot \nabla]u(x,t) = f(x,t)\ , \tag{1.1}$$

$$u(x,t) \in H^l \cap H^{l+1}_{m\ell}(\gamma) \quad \textit{near } \{t = 0\}\ , \tag{1.2}$$

with

$$\gamma = \{(x,t,\xi,\omega), (x,t) \in \Omega, (\xi,\omega) \in \mathbf{R}^{n+1}, |\xi'|^2 = d^2\omega^2 \textit{ and } \xi_n = 0\}\ ,$$

where Ω is a compact subset of $\{(x,t) \in \mathbf{R}^{n+1}, |t|, |x_n| \le \epsilon_0\}$ for small $\epsilon_0 > 0$ and

$$\Gamma = \textit{a conic neighborhood of } \gamma\ .$$

Assume that

(i) $u \in H^{l-1} \cap H^l_{m\ell}(\Gamma)$, $1 \le l \le s$;

(ii) $d(x) \in H^s \cap H^{l+1}_{m\ell}(K)$ *and* $\nabla\sigma(x) \in H^{s-1} \cap H^l_{m\ell}(K)$, $\Pi :\ \Gamma \subset T^*(\mathbf{R}^{n+1}) \to K \subset T^*(\mathbf{R}^n)$ *is the projection map*;

(iii) $f \in H^{l-1} \cap H^l_{m\ell}(\Gamma)$.

Then

$$u\,|_{x_n=0} \in H^l_{\mathrm{loc}}\ ,$$

Trace regularity results are of great importance in diverse fields, such as control theory, boundary and initial boundary value problems of partial differential equations (PDEs), and in particular inverse problems. An interesting example arises from seismic imaging, where one wants to determine the mechanical properties of the earth through analyzing the seismogram measured on the surface of earth, i.e., the solution on a time-like surface. Clearly, modeling this process requires the most precise information about the time-like trace.

In [2], we proved a trace theorem for general linear PDEs with smooth variable coefficients, applying the Hörmander–Nirenberg pseudodifferential cutoff technique and the method of energy estimates. Our theorem indicates that the difficulty mentioned above may be resolved by imposing more smoothness against grazing ray directions. We then studied in [3] the wave equation with nonsmooth coefficients only at lower-order terms. Compared to the general case in [2], a much simpler pseudodifferential cutoff was available, which together with a result on propagation of singularities for strictly hyperbolic $\psi.d.o.$ equations with nonsmooth coefficients at lower order terms enabled us to prove a trace theorem.

The proof of Theorem 1.1 basically follows the framework developed in [3]. However, because of the presence of nonsmooth coefficients in the principal part, additional technical difficulties will necessarily take place. For example, the result on propagation of singularities which is essential in our proof becomes much harder to obtain. When the principal part has smooth or constant coefficients, as in [3], a generalized Rauch's lemma and a commutator lemma allowed us to extend a Beals–Reed theorem (the BRI, Theorem 1 in [7]) on propagation of singularities to a more suitable setting

for linear problems. However, the same idea breaks down when the principal part becomes nonsmooth. To cure the difficulty, a refined approach was introduced by Beals and Reed in [8], where nonsmooth symbol classes and the corresponding calculus were developed. Their approach follows the general outline of Hörmander's proof of the theorem on propagation of singularities for PDEs with smooth coefficients [11], but of course the nonsmoothness of the principal part requires some very delicate commutator estimates. The key step was to establish a sharp Gårding's inequality which allowed them to reduce the microlocal problem to a local one. They then proved a theorem on propagation of singularities for $\psi.d.o$ equations with nonsmooth symbols (the BRII, Theorem 3.2 in [8]). In this paper, we prove an extended version of the Beals–Reed theorem (the BRII). Our theorem assures that under similar hypotheses some lower-order microlocal regularity of the solution, for instance, H^s_{ml}-regularity (for $s \geq 0$), will also propagate along the null bicharacteristics. We also treat the case where coefficients may depend only on some of the variables, while we show that a better result can be expected. We emphasize that this is particularly important since the coefficients in our model are time independent!

We expect that the calculus of nonsmooth symbols developed in [8] and this work will find other applications. Working on PDEs with nonsmooth coefficients, one would often encounter similar commutator estimates.

Recently, studies on trace regularity have received much attention. For solutions of the multidimensional wave equation with constant coefficients, Symes proved in [22], by using the method of geometric optics, that with finitely energy initial data compactly supported away from the boundary, the trace is as regular as the solution in the interior. Similar results were obtained independently by Lasiecka and Triggiani [16] based on an application of the classical Laplace–Fourier transform. They also established (see [17]) a series of sharp global trace regularity results for second order hyperbolic equations with smooth coefficients and Dirichlet or Neumann boundary conditions. More recently, by using a microlocal cutoff technique, Lasiecka and Triggiani in [18] obtained an optimal tangential trace regularity theorem for the wave equation with smooth coefficients in a bounded domain. For a wave equation, Kim [13] proved a trace regularity result by employing a microlocal method, which further allowed him to study the regularity of some boundary controls. Our approach differs from that mentioned above in several aspects: The data in this work need not have compact supports, and the coefficients in the model may be nonsmooth; furthermore, our method relies heavily on various results on propagation of singularities. For more general geometry and functions, Joly studied in [15] some trace regularity results in the case of anisotropic Sobolev spaces in two dimensions where the domain can be any arbitrary open set of $\mathbf{R}^2$.

The outline of the paper is as follows: In the next section, we introduce nonsmooth symbol classes which are more flexible than those defined in [8] in the sense that the coefficients may depend only on part of the variables. We then develop a calculus for this class of nonsmooth symbols which handles a more general class of functions with less regularity than in [8]. We establish a Gårding type inequality. Finally, we present a theorem on propagation of singularities. Section 3 is devoted to the proof of our main theorem. This is done by a combination of microlocal cutoff, the "fattening" lemma, the method of energy estimates, and the theorem on propagation of singularities.

Throughout, the reader is assumed to be familiar with the basic calculus of *pseudodifferential operators* ($\psi.d.o.$) as presented in [24] and [19]. A classical $\psi.d.o.$ P of

order m is denoted as $P \in OPS^m$ with its symbol $p \in S^m$. $ES(P)$ stands for *the essential support* of operator P. H^s is the standard L^2-type Sobolev space and H^s_{loc} means a local Sobolev space. $\langle\xi\rangle$ means $(1+|\xi|^2)^{1/2}$. The Fourier transform of a distribution u is expressed as $\hat{u}$. Usually, the constant from the Fourier transform is assumed to be absorbed by the integral. For simplicity, C serves as a generalized positive constant the precise value of which is not needed. The characteristic function of a set Γ is represented by χ_Γ.

We conclude the introduction by a brief discussion about nonlinear wave equations. The approach of this work may be adopted nicely in the trace regularity study of a class of nonlinear wave equations. In fact, various results on propagation of singularities for semilinear and quasi-linear wave equations have been developed by many people, in particular by Rauch [20], Bony [9], Beals and Reed [7], [8], Beals [4], [5], and Chemin [10]. But compared to the linear case, much higher overall smoothness of both the solution and coefficients would be required. Along this direction, some progress has been made in Bao [1].

2. Propagation of microlocal regularity. We introduce some basic concepts, related material may be found in Beals and Reed [8]. General theoretical aspects as well as applications of nonsmooth microlocal analysis may be found in Rauch [20], Beals [6], and Taylor [25] and references therein.

DEFINITION 2.1. *A distribution u is said to be in $H^s \cap H^r_{m\ell}(x_0,\xi_0)$ if there exist $\phi(x) \in C_0^\infty(\mathbf{R}^n)$ with $\phi(x_0) \neq 0$ and a conic neighborhood $\gamma \subset \mathbf{R}^n\backslash\{0\}$ of ξ_0 such that*

$$\langle\xi\rangle^s(\phi u)^\wedge(\xi) \in L^2(\mathbf{R}^n) \quad \text{and} \quad \langle\xi\rangle^r\chi_\gamma(\xi)(\phi u)^\wedge(\xi) \in L^2(\mathbf{R}^n)\,.$$

Next, we define a nonsmooth symbol class, which follows [8] with one main difference, that is, the present symbol class includes symbols that depend only on part of the variables. An interesting special case deals with PDEs with coefficients depending on part of the variables. For convenience, we denote $x = (x^1, x^2) \in \mathbf{R}^{n_0} \times \mathbf{R}^{n-n_0}$, $1 \leq n_0 \leq n$. The frequency variables corresponding to $x = (x^1, x^2)$ are denoted by $\xi = (\xi^1, \xi^2) \in \mathbf{R}^{n_0} \times \mathbf{R}^{n-n_0}$.

DEFINITION 2.2. *$S^{m;s,r}_{n_0}(K)$ is defined to be the collection of symbols $a(x^1,\xi)$, smooth in ξ, such that*

$$a(x^1,\xi)/\langle\xi\rangle^m \in H^s \cap H^r_{m\ell}(K)$$

as a function of x^1 uniformly in ξ.

In other words, for each $(x^1_0,\xi_0) \in K$, there exist a function $\psi(x^1) \in C_0^\infty(\mathbf{R}^{n_0})$ with $\psi(x^1_0) \neq 0$ and a conic neighborhood γ of ξ_0 such that

$$\langle\zeta\rangle^s\mathcal{F}(\psi a)(\zeta,\xi)/\langle\xi\rangle^m \in L^2(d\zeta)$$

and

$$\langle\zeta\rangle^r\chi_\gamma(\zeta)\mathcal{F}(\psi a)(\zeta,\xi)/\langle\xi\rangle^m \in L^2(d\zeta)$$

with norms independent of ζ.

In particular, when $n = n_0$, the symbol class becomes the one introduced in [8]. In that case, we denote

$$S^{m;s,r}(K) = S^{m;s,r}_n(K)\,.$$

More generally, dealing with operators with lower order terms, we also need the following definition.

DEFINITION 2.3. *For $k \geq 0$ an integer, $S^{m+k;s+k,r+k}_{st,n_0}(K)$ consists of symbols $a(x^1,\xi)$ of the form*

$$a_{m+k}(x^1,\xi) + a_{m+k-1}(x^1,\xi) + \cdots + a_m(x^1,\xi),$$

where $a_m(x^1,\xi) \in S^{m;s,r}(K)$ and for $0 < j \leq k$

$$a_{m+j}(x^1,\xi) = \sum_l a_{j,l}(x^1) P_{m+j,l}(x^1,\xi) \in S^{m;s,r}(K)$$

with $a_{j,l}(x^1) \in H^{s+j} \cap H^{r+j}_{m\ell}(K)$ and $P_{m+j,l}(x^1,\xi) \in S^{m+j}_{1,0}$.

It is easy to observe that

$$S^{m+k;s+k,r+k}_{st,n_0}(K) \subset S^{m+k;s+k,r+k}_{n_0}(K) + \cdots + S^{m;s,r}_{n_0}(K) \subset S^{m+k;s,r}_{n_0}(K)\,.$$

We are now ready to state a linear theorem on propagation of singularities for ψ.d.o. equations with nonsmooth coefficients. Again recall that $x = (x^1,x^2) \in \mathbf{R}^{n_0} \times \mathbf{R}^{n-n_0}$, and $\xi = (\xi^1,\xi^2)$ is the corresponding frequency variable.

THEOREM 2.1. *Let $n_0/2 < s$, $0 \leq l \leq s, q$ and $q < l + s - n_0/2$ $(1 \leq n_0 \leq n)$, and*

$$a(x^1,\xi) \in S^{m+2;s+2,q+2}_{st,n_0}(K)$$

is of real principal type with real principal symbol a_{m+2} which is homogeneous of degree $m+2$ in ξ, and that γ is the null bicharacteristic through the characteristic point (x_0,ξ_0). $K = \Pi\gamma$ with $\Pi : T^*(\mathbf{R}^n) \to T^*(\mathbf{R}^{n_0})$ the projection map. Assume that

(i) $v \in H^{l+m} \cap H^{q+m}_{m\ell}(\gamma)$,
(ii) $v \in H^{l+m} \cap H^{q+m+\epsilon}_{m\ell}(x_0,\xi_0)$ *for some* $0 \leq \epsilon \leq 1$,
(iii) $f \in H^l \cap H^q_{m\ell}(\gamma)$,

and that

$$a(x^1,D_x)v = f.$$

Then

$$v \in H^{q+m+\epsilon}_{m\ell}(\gamma).$$

That is, if the solution v has improved microlocal regularity at a point on a null bicharacteristic, the improved regularity will stay for the rest of the null bicharacteristic.

Remarks on Theorem 2.1. In the case that $l = s$ and $n_0 = n$, we speak of the original Beals–Reed theorem (the BRII). Compared to the BRII, Theorem 2.1 assures that weaker regularity of the solution may also be propagated along null bicharacteristics, and a better regularity result can be achieved when the coefficients depend only on some of the variables.

It seems that the regularity hypotheses on the coefficients and right-hand side cannot be improved much. However, the sharpness questions on this theorem and other theorems on propagation of singularities (the BRI, the BRII, and Theorem 2.1 in [3]) remain open. Needless to say, these questions are extremely important and worthwhile to pursue in the future.

The most precise information about the propagation of singularities may be obtained in the case of one-space dimension, where the wave operator can be factored into products of two simple first-order differential operators. Roughly speaking, the improved microlocal regularity is then propagated along null bicharacteristics with very few restriction on the order of smoothness. The best reference for one-dimensional hyperbolic problems is Rauch and Reed [21].

2.1. About the proof of Theorem 2.1. Theorem 2.1 may be proved by following the general scheme of the proof of the BRII in [8] with some necessary modifications. The main ingredients of their proof are as follows:

- The development of a calculus for the nonsmooth symbol class
- Construction of an appropriate microlocal cutoff b_0 by essentially following Bony's construction in [9].
- Proof of a generalized Gårding's inequality.

The key step was to develop the calculus of the nonsmooth symbol class, which led to a systematic way of handling commutators that involve either nonsmooth functions or nonsmooth symbols ($\psi.d.o.$ with nonsmooth coefficients). It was shown that the nonsmooth symbol class developed by Beals and Reed preserves important continuity properties of smooth $\psi.d.o.$

We point out that when the differential operator of a differential equation has constant coefficients or is with a smooth principal part, the microlocal cutoff may be constructed in a much simpler fashion. In fact, following Hörmander's first proof of the theorem on propagation of singularities in [12] (see also Nirenberg [19]) there exists a $b_0 \in OPS^0$ such that $[P_m, b_0]$ is of order $m-2$ instead of $m-1$, if P_m is the principal part of an mth order partial differential operator with usual assumption on P_m. Having constructed this operator b_0, one may reduce the microlocal regularity problem to a local one by acting b_0 to both sides of the PDE.

For simplicity, we shall skip the formal proof of Theorem 2.1 with the understanding that the proof can be formally done by modifying the proof of the BRII ([8], pp. 174–177). In order to do so, a calculus of general nonsmooth symbols must be developed and also a new Gårding's inequality must be established. Through this process, Theorem 2.2 will play a crucial role since it characterizes the fundamental Sobolev continuous properties of the class of nonsmooth symbols. The rest of this section is devoted to establish the necessary results for completing the proof of Theorem 2.1.

2.2. Sobolev continuity of nonsmooth symbols. We begin with some inequalities that will be used frequently.

PROPOSITION 2.1. *Let*

$$C_g^2 = \sup_{\xi,\eta} \int |g(\zeta,\xi,\eta)|^2 d\zeta$$

and

$$C_K^2 = \sup_{\zeta,\eta} \int |K(\zeta,\xi,\eta)|^2 d\xi \,.$$

For $h \in L^2$, define

$$(Th)(\zeta,\eta) = \int K(\zeta,\xi,\eta) g(\zeta-\xi,\xi,\eta) h(\xi,\eta) d\xi \,.$$

Then

$$||Th|| \le C_g C_K ||h|| \,.$$

Proof. Consider that for any $f(\zeta,\eta) \in L^2$ and $||f|| \le 1$,

$$|(Th,f)| = \left| \int \int f(Th) d\zeta d\eta \right|$$

$$= \left| \int\int\int K(\zeta,\xi,\eta) g(\zeta-\xi,\xi,\eta) f(\zeta,\eta) h(\xi,\eta) d\xi d\zeta d\eta \right| .$$

Interchanging the sequence of integrations and by Cauchy–Schwarz's inequality, we have

$$\begin{aligned} |(Th,f)| &\le \int\int d\xi d\eta |h(\xi,\eta)| \left[\int d\zeta |K(\zeta,\xi,\eta) f(\zeta,\eta)|^2\right]^{1/2} \cdot \left[\int d\zeta |g(\zeta-\xi,\xi,\eta)|^2\right]^{1/2} \\ &\le C_g ||h|| \cdot \left[\int\int\int d\xi d\eta d\zeta |K(\zeta,\xi,\eta) f(\zeta,\eta)|^2\right]^{1/2} \\ &\le C_g C_K \, ||h|| \cdot ||f||, \end{aligned}$$

hence the conclusion follows. □

Similarly, one can prove the following result.

PROPOSITION 2.2. *Denote* $\zeta = (\zeta^1, \zeta^2)$ *and* $\eta = (\eta^1, \eta^2)$, *where* ζ^1, η^1 *and* ξ *are of the same dimension. Let*

$$C_h^2 = \sup_{\zeta} \int d\xi |h(\xi,\zeta)|^2$$

$$C_g^2 = \sup_{\zeta,\xi} \int d\eta |g(\eta,\zeta,\xi)|^2$$

and

$$C_G^2 = \sup_{\eta} \int\int d\xi d\zeta |G(\xi,\eta,\zeta)|^2 .$$

For $v \in L^2$, *define*

$$T(v)(\eta) = \int\int d\xi d\zeta G(\xi,\eta,\zeta) h(\xi,\zeta) g(\eta^1-\zeta^1-\xi, \eta^2-\zeta^2, \zeta, \xi) v(\zeta) .$$

Then

$$||T(v)|| \le C_h C_g C_G ||v|| .$$

Proof. For any $f \in L^2(d\eta)$ with $||f|| \le 1$, consider

$$\begin{aligned} |(T(v),f)| &\le \int d\zeta |v(\zeta)| \int\int d\xi d\eta |h(\xi,\zeta) g(\eta^1-\zeta^1-\xi, \eta^2-\zeta^2, \zeta, \xi)| |G(\xi,\eta,\zeta) f(\eta)| \\ &\le \int d\zeta |v(\zeta)| \left[\int d\xi |h|^2 \int d\eta |g|^2\right]^{1/2} \left[\int\int d\xi d\eta |G|^2 |f|^2\right]^{1/2} \\ &\le C_g C_h \int d\zeta |v(\zeta)| \left[\int\int d\xi d\eta |G|^2 |f|^2\right]^{1/2} \\ &\le C_g C_h C_G ||v|| \, ||f||. \end{aligned}$$ □

PROPOSITION 2.3. *Assume that* K' *is a closed cone which is strictly contained in an open cone* K. *If* $\xi \in K'$, $\eta \in K^C$, *where* K^C *is the complement of* K, *then*

(1) $|\xi - \eta| \ge C_1 |\xi|$, $C_1 > 0$;

(2) *if* $|\xi| \ge C_0 > 0$, *then* $\langle \xi - \eta \rangle \ge C \langle \xi \rangle$.

The following theorem establishes continuity properties of the nonsmooth symbol class. Let $\Pi_1 : (x^1, x^2, \xi^1, \xi^2) \to (x^1, \xi^1, \xi^2)$.

THEOREM 2.2. *Let $n_0/2 < s$, $0 \le l \le s, q$, and $q < l + s - n_0/2$ $(1 \le n_0 \le n)$. Suppose that $a(x^1, \xi) \in S^{m;s,q}_{n_0}(\gamma)$, $0 \le m \le s, l$, $\gamma = \Pi_1\Gamma$. Then*

$$a(x^1, D_x)$$

is a bounded operator from $H^l \cap H^q_{m\ell}(\Gamma)$ to $H^{l-m} \cap H^{q-m}_{m\ell}(\Gamma)$.

Remarks. We list some of the interesting special cases of Theorem 2.2.

- If $n_0 = n$ and $l = s$, then Theorem 2.2 becomes an earlier result of Beals and Reed (Theorem 1.3 in [8]). In addition, the case where $a(x, \xi) = a(x^1, \xi)$ gives rise to Rauch's lemma [20].
- The case $a(x^1, \xi) = a(x)$ is the generalized Rauch's lemma proved in [3] (Lemma 2.3).

Proof. W.l.o.g., we assume that $s < q$, some obvious modification will yield the conclusion for $s \ge q$. For $u \in H^l \cap H^q_{m\ell}(\Gamma)$, we may assume u and $a(x^1, \xi)$ have compact supports in x near x_0, and x^1 near x^1_0, respectively, where $(x_0, \xi_0) \in \Gamma$. Then

$$a(x^1, D_x)u(x) = \int\int e^{ix\cdot\xi} a(x^1, \xi)\hat{u}(\xi)d\xi$$

and

$$\widehat{(au)}(\eta) = \int \hat{a}(\eta^1 - \xi^1, \xi^1, \eta^2)\hat{u}(\xi^1, \eta^2)d\xi^1 \, .$$

Next since $a \in S^{m;s,r}_{n_0}(\gamma)$,

$$\hat{a}(\zeta, \xi) = f(\zeta, \xi)\langle\xi\rangle^m/\langle\zeta\rangle^s$$

where $\sup_\xi \int f^2(\zeta, \xi)d\zeta < \infty$.

Define $v(\xi) = \langle\xi\rangle^l\hat{u}(\xi) \in L^2$. Thus

$$\langle\eta\rangle^{l-m}\widehat{(au)}(\eta) = \int K(\xi^1, \eta)f(\eta^1 - \xi^1, \xi^1, \eta^2)v(\xi^1, \eta^2)d\xi^1$$

with

$$K(\xi^1, \eta) = \langle\eta\rangle^{l-m}/\langle\eta^1 - \xi^1\rangle^s\langle\xi^1, \eta^2\rangle^{l-m} \, .$$

Therefore from Proposition 2.1, it suffices to show that

$$\sup_\eta \int |K(\xi^1, \eta)|^2 d\xi^1 < \infty$$

which follows immediately from Hölder's inequality and the assumption that $s > n_0/2$ and $l \ge m$.

Let θ be a conic neighborhood of ξ_0 such that $u \in H^l \cap H^q_{m\ell}(\theta)$, and $a(x^1, \xi) \in S^{m;s,r}_{n_0}(\sigma)$ where $\sigma = \Pi\theta$ the projection of θ onto the ξ^1-space. Let $\theta' \subset\subset \theta$, be a strictly smaller conic neighborhood of ξ_0 and σ' is the projection of θ'. We must show that

$$\chi_{\theta'}(\eta)\langle\eta\rangle^{q-m}\widehat{(au)}(\eta) \in L^2 \, .$$

Write

$$\hat{a}(\zeta,\xi) = \frac{\chi_\sigma(\zeta) f_1(\zeta,\xi)\langle\xi\rangle^m}{\langle\zeta\rangle^q} + \frac{\chi_{\sigma^C}(\zeta) f_2(\zeta,\xi)\langle\xi\rangle^m}{\langle\zeta\rangle^s}$$
$$\hat{u}(\xi) = \frac{\chi_\theta(\xi) v_1(\xi)}{\langle\xi\rangle^q} + \frac{\chi_{\theta^C}(\xi) v_2(\xi)}{\langle\xi\rangle^l},$$

where $f_i \in L^2(d\zeta)$ uniformly in ξ and $v_i \in L^2$. Then

$$\chi_{\theta'}(\eta)\langle\eta\rangle^{q-m}\hat{u} = \sum_{i,j=1,2} \int K_{ij}(\xi^1,\eta) f_i(\eta^1-\xi^1,\xi^1,\eta^2) v_i(\xi^1,\eta^2) d\xi^1,$$

where

$$K_{11}(\xi^1,\eta) = \frac{\chi_{\theta'}(\eta)\chi_\sigma(\eta^1-\xi^1)\chi_\theta(\xi^1,\eta^2)\langle\xi^1,\eta^2\rangle^m\langle\eta\rangle^{q-m}}{\langle\eta^1-\xi^1\rangle^q\langle\xi^1,\eta^2\rangle^q}$$
$$K_{12}(\xi^1,\eta) = \frac{\chi_{\theta'}(\eta)\chi_\sigma(\eta^1-\xi^1)\chi_{\theta^C}(\xi^1,\eta^2)\langle\xi^1,\eta^2\rangle^m\langle\eta\rangle^{q-m}}{\langle\eta^1-\xi^1\rangle^q\langle\xi^1,\eta^2\rangle^l}$$
$$K_{21}(\xi^1,\eta) = \frac{\chi_{\theta'}(\eta)\chi_{\sigma^C}(\eta^1-\xi^1)\chi_\theta(\xi^1,\eta^2)\langle\xi^1,\eta^2\rangle^m\langle\eta\rangle^{q-m}}{\langle\eta^1-\xi^1\rangle^s\langle\xi^1,\eta^2\rangle^q}$$
$$K_{22}(\xi^1,\eta) = \frac{\chi_{\theta'}(\eta)\chi_{\sigma^C}(\eta^1-\xi^1)\chi_{\theta^C}(\xi^1,\eta^2)\langle\xi^1,\eta^2\rangle^m\langle\eta\rangle^{q-m}}{\langle\eta^1-\xi^1\rangle^s\langle\xi^1,\eta^2\rangle^l}.$$

Once again, by Proposition 2.1, it suffices to show that

$$\sup_\eta \int |K_{ij}(\xi^1,\eta)|^2 d\xi^1 < \infty.$$

We now estimate these kernels separately.

On support K_{11}, $\eta \in \theta'$, $\eta^1-\xi^1 \in \sigma$, $(\xi^1,\eta^2) \in \theta$. Then

$$|K_{11}| \le \frac{\langle\eta\rangle^{q-m}}{\langle\eta^1-\xi^1\rangle^q\langle\xi^1,\eta^2\rangle^{q-m}} \in L^2(d\xi^1).$$

On support K_{12}, $\eta \in \theta'$, $\eta^1-\xi^1 \in \sigma$, and $(\xi^1,\eta^2) \in \theta^C$. Then $\langle\eta^1-\xi^1\rangle \ge C\langle\eta\rangle$, hence

$$|K_{12}| \le \frac{1}{\langle\eta^1-\xi^1\rangle^m\langle\xi^1,\eta^2\rangle^{l-m}} \in L^2(d\xi^1)$$

since $l > n_0/2$.

On support K_{21}, $\eta \in \theta'$, $\eta^1-\xi^1 \in \sigma^C$, and $(\xi^1,\eta^2) \in \theta$, therefore $\langle\xi^1,\eta^2\rangle \ge C\langle\eta\rangle$. It follows that

$$|K_{21}| \le \frac{C}{\langle\eta^1-\xi^1\rangle^s} \in L^2(d\xi^1).$$

On support K_{22}, $\eta \in \theta'$, $\eta^1-\xi^1 \in \sigma^C$, and $(\xi^1,\eta^2) \in \theta^C$, therefore $\langle\xi^1,\eta^2\rangle \ge C\langle\eta\rangle$. It follows that

$$|K_{22}| \le \frac{C}{\langle\eta^1-\xi^1\rangle^{s+l-q}} \in L^2(d\xi^1)$$

because of the assumption $s+l-q > n_0/2$. □

2.3. Calculus of nonsmooth symbol class. With the presence of nonsmooth coefficients, one always needs to study various algebraic properties, such as products of two nonsmooth functions, the action of a nonsmooth symbol ($\psi.d.o.$) to a function, and commutators between smooth (or nonsmooth) $\psi.d.o.$s. To serve the proof of Theorem 2.1, we shall study compositions of an operator that has nonsmooth symbol with a smooth operator from left. The proof of Theorem 2.1 also requires a Gårding's inequality.

DEFINITION 2.4. *For $\Gamma \subset T^*(\mathbf{R}^n)$, $S_{bd}^{m;s,q,l,r}(\Gamma)$ is a collection of symbols that are in $S^{m;s,q}(\Gamma)$ and define bounded maps from*

$$H^l \cap H^r_{m\ell}(\Gamma) \to H^{l-m} \cap H^{r-m}_{m\ell}(\Gamma) .$$

LEMMA 2.1. *Let $n_0/2 < s$, $0 \le l \le s$, $1 \le n_0 \le n$, $s \le s_b$, $n/2 < s_b$, and $q < \min\{s + l - n_0/2, s_b + s - n/2\}$. Let $k \ge 0$ be an integer, and for $|\alpha| \le k$*

$$a(x^1, \xi) \in S^{m;s+k,q+k}_{n_0}(\gamma) , \quad \partial^\alpha_\xi b(x, \xi) \in S^{k-|\alpha|;s_b,q}(\Gamma) , \text{ and } \gamma = \Pi_1\Gamma .$$

Then $b \circ a(x, \xi)$, the symbol of operator $b(x, D)a(x^1, D)$, satisfies

$$\left\{ b \circ a(x, \xi) - \sum_{|\alpha|<k} \frac{1}{\alpha!} \partial^\alpha_\xi b D^\alpha_{x^1} a(x^1, \xi) \right\} \in S^{m;s,q,l,q}_{bd}(\Gamma) .$$

Proof.

$$(b \circ a)(x, D_x)u(x) = \int e^{ix\cdot\xi} b(x, \xi)\widehat{(au)}(\xi) d\xi.$$

It follows from

$$\widehat{(au)}(\xi) = \int \hat{a}(\xi^1 - \zeta^1, \zeta^1, \xi^2)\hat{u}(\zeta^1, \xi^2) d\zeta^1$$

that

$$(b \circ a)(x, D_x)u(x) = \int\int e^{ix\cdot\xi} b(x, \xi)\hat{a}(\xi^1 - \zeta^1, \zeta^1, \xi^2)\hat{u}(\zeta^1, \xi^2) d\zeta^1 d\xi$$

and

$$\widehat{(b \circ au)}(\eta) = \int\int \hat{b}(\eta - \xi, \xi)\hat{a}(\xi^1 - \zeta^1, \zeta^1, \xi^2)\hat{u}(\zeta^1, \xi^2) d\zeta^1 d\xi .$$

Substituting ξ^1 and ζ^2 by $\xi^1 + \zeta^1$ and ξ^2, we have

$$\widehat{(b \circ au)}(\eta) = \int\int \hat{b}(\eta^1 - \xi^1 - \zeta^1, \eta^2 - \zeta^2, \xi^1 + \zeta^1, \zeta^2)\hat{a}(\xi^1, \zeta^1, \zeta^2)\hat{u}(\zeta) d\zeta d\xi^1 .$$

On the other hand,

$$(\partial^\alpha_\xi b D^\alpha_{x_1} a)(x, D_x)u(x) = \int e^{ix\cdot\zeta} (\partial^\alpha_\zeta b D^\alpha_{x_1} a)(x, \zeta)\hat{u}(\zeta) d\zeta$$

and

$$\begin{aligned} \widehat{(\partial^\alpha_\xi b D^\alpha_{x_1} au)}(\eta) &= \int\int e^{ix\cdot(\zeta-\eta)} (\partial^\alpha_\zeta b D^\alpha_{x_1} a)(x, \zeta)\hat{u}(\zeta) d\zeta dx \\ &= \int\int \partial^\alpha_\zeta \hat{b}(\eta^1 - \zeta^1 - \xi^1, \eta^2 - \zeta^2, \zeta)(\xi^1)^\alpha \hat{a}(\xi^1, \zeta)\hat{u}(\zeta) d\zeta d\xi^1 . \end{aligned}$$

Now, defining

$$r(x,D) = b \circ a(x,D) - \sum_{|\alpha|<k} \frac{1}{\alpha!}(\partial_\xi^\alpha b D_{x_1}^\alpha a)(x,D)$$

we then get by Taylor's theorem

(2.1)

$$(\widehat{ru})(\eta) = \sum_{|\alpha|=k} C_\alpha \int\int \left\{ \int_0^1 (1-t)^{k-1} (\partial_\zeta^\alpha \hat{b})(\eta^1 - \zeta^1 - \xi^1, \eta^2 - \zeta^2, \zeta^1 + t\xi^1, \zeta^2) dt \right\} \times (\xi^1)^\alpha \hat{a}(\xi^1, \zeta) \hat{u}(\zeta) d\xi^1 d\zeta$$

for the case $k = 0$, the term in braces should be replaced by $\hat{b}(\eta^1 - \zeta^1 - \xi^1, \eta^2 - \zeta^2, \zeta^1 + t\xi^1, \zeta^2)$.

Therefore

$$(\widehat{ru})(\eta) = \int \hat{r}(\eta - \zeta, \zeta) \hat{u}(\zeta) d\zeta \,,$$

where

$$\hat{r}(\eta,\zeta) = \begin{cases} \displaystyle\int \sum_{|\alpha|=k} C_\alpha \int \left\{ \int_0^1 (1-t)^{k-1} (\partial_\zeta^\alpha \hat{b})(\eta^1 - \xi^1, \eta^2, \zeta^1 + t\xi^1, \zeta^2) dt \right\} \\ \qquad \times (D_{x^1}^\alpha a)\widehat{\ }(\xi^1, \zeta) d\xi^1 \quad \text{if } k \neq 0 \\ \displaystyle\int \hat{b}(\eta^1 - \xi^1, \eta^2, \xi^1 + \zeta^1, \zeta^2) \hat{a}(\xi^1, \zeta) d\xi^1 \quad \text{if } k = 0. \end{cases}$$

Let $K' \subset\subset K$ be a small conic neighborhood of ξ_0, $(x_0, \xi_0) \in \Gamma$, with K sufficiently small so that estimates of Definition 2.2 hold for $\partial_\zeta^\alpha b$ on K and a on $\theta = \Pi K$ (Π the projection map from $\mathbf{R}^n$ to $\mathbf{R}^{n_0}$). In order to show

$$\left\{ b \circ a(x,\xi) - \sum_{|\alpha|<k} \frac{1}{\alpha!} \partial_\xi^\alpha b D_{x^1}^\alpha a(x^1, \xi) \right\} \in S^{m;s,q}(\Gamma)$$

it suffices to prove

$$\langle \eta \rangle^s \frac{\hat{r}(\eta,\zeta)}{\langle \zeta \rangle^m} \in L^2(d\eta) \text{ and } \langle \eta \rangle^q \frac{\chi_{K'}(\eta) \widehat{\ }(r)(\eta,\zeta)}{\langle \zeta \rangle^m} \in L^2(d\eta).$$

This may be proved similarly as in the proof of Theorem 2.2 by treating ζ as a parameter, hence we shall omit it.

We next show that $r(x,D)$ is a bounded operator from $H^l \cap H_{m\ell}^q(\Gamma)$ to $H^{l-m} \cap H_{m\ell}^{q-m}(\Gamma)$.

Let us begin by proving the local part of the conclusion. W.l.o.g., only the case $k \neq 0$ is considered here. For $|\alpha| = k$, let

$$\langle \xi^1 \rangle^\alpha \hat{a}(\xi^1, \zeta) = \frac{h(\xi^1, \zeta) \langle \zeta \rangle^m}{\langle \xi^1 \rangle^s} \text{ and } \hat{u}(\zeta) = v(\zeta)/\langle \zeta \rangle^l,$$

where $h(\xi^1,\zeta) \in L^2(d\xi^1)$ uniformly in ζ and $v \in L^2$. The integral kernel in the expression (2.1) can be written

$$g(\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2,\zeta,\xi^1)/\langle\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2\rangle^{s_b},$$

where $g(\eta,\zeta,\xi^1) \in L^2(d\eta)$ uniformly in ζ and ξ^1.

Thus

$$\langle\eta\rangle^{l-m}\widehat{(ru)}(\eta) = \sum_{|\alpha|=k} C_\alpha \int\int K(\xi^1,\eta,\zeta)g(\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2,\zeta,\xi^1)h(\xi^1,\zeta)v(\zeta)d\xi^1 d\zeta$$

with

$$K(\xi^1,\eta,\zeta) = \frac{\langle\eta\rangle^{l-m}}{\langle\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2\rangle^{s_b}\langle\xi^1\rangle^s\langle\zeta\rangle^{l-m}}\,.$$

Because of Proposition 2.2, we only need to show

$$K \in L^2(d\xi^1 d\zeta) \quad \text{uniformly in } \eta\,.$$

We separate it into several pieces:

If $|\eta|/3 \le |\zeta|$, then

$$|K| \le \frac{C}{\langle\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2\rangle^{s_b}\langle\xi^1\rangle^s};$$

hence $K \in L^2(d\xi^1 d\zeta)$, since $s > n_0/2$ and $s_b > n/2$.

In the region where $|\zeta| < |\eta|/3 \le |(\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2)|$, then

$$\begin{aligned}|K| &\le \frac{C}{\langle\xi^1-\zeta^1-\xi^1,\eta^2-\zeta^2\rangle^{s_b-l+m}\langle\xi^1\rangle^s\langle\zeta\rangle^{l-m}}\\ &\le \frac{C}{\langle\xi^1\rangle^s\langle\zeta\rangle^{s_b}}.\end{aligned}$$

In the region, $|\zeta| < |\eta|/3$ and $|(\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2)| < |\eta|/3$. Since

$$\eta = \zeta + (\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2) + (\xi^1,0)$$

we have

$$|\eta| \le |\zeta| + |(\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2)| + |\xi^1|$$

or $|\eta|/3 \le |\xi^1|$. Thus

$$\begin{aligned}|K| &\le \frac{C}{\langle\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2\rangle^{s_b}\langle\xi^1\rangle^{s-l+m}\langle\zeta\rangle^{l-m}}\\ &\le \frac{C}{\langle\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2\rangle^{s_b}\langle\zeta\rangle^s},\end{aligned}$$

i.e., $K \in L^2(d\xi^1 d\zeta)$.

We next deal with the microlocal part. Let $\theta' \subset\subset \theta$, then for $|\alpha| = k$, write

$$\langle\xi^1\rangle^\alpha \hat{a}(\xi^1,\zeta) = \frac{\chi_{\theta'}h_1(\xi^1,\zeta)\langle\zeta\rangle^m}{\langle\xi'\rangle^q} + \frac{\chi_{\theta^C}(\xi^1)h_2(\xi^1,\zeta)\langle\zeta\rangle^m}{\langle\xi^1\rangle^s},$$

where $h_i(\xi^1,\zeta) \in L^2(d\xi^1)$ uniformly in ζ. Let

$$\hat{u}(\zeta) = \frac{\chi_{K'}(\zeta)v_1(\zeta)}{\langle\zeta\rangle^q} + \frac{\chi_{K^C}(\zeta)v_2(\zeta)}{\langle\zeta\rangle^l}$$

with v_1 and $v_2 \in L^2$.

The integral kernel of (2.1) can be written as

$$\frac{\chi_{K'}(\eta^1-\zeta^1-\xi^1,\eta^2-\xi^2)g_1(\eta^1-\zeta^1-\xi^1,\eta^2-\xi^2,\zeta,\xi^1)}{\langle\eta^1-\zeta^1-\xi^1,\eta^2-\xi^2\rangle^r} + \frac{\chi_{K^C}(\eta^1-\zeta^1-\xi^1,\eta^2-\xi^2)g_2(\eta^1-\zeta^1-\xi^1,\eta^2-\xi^2,\zeta,\xi^1)}{\langle\eta^1-\zeta^1-\xi^1,\eta^2-\xi^2\rangle^{s_b}},$$

where $g_i(\eta,\zeta,\xi^1) \in L^2(d\eta)$ uniformly in ξ^1 and ζ.

Therefore, we have the following close form:

$$\begin{aligned}&\chi_{K'}(\eta)\langle\eta\rangle^{q-m}\widehat{(ru)}(\eta)\\ &= \sum_{|\alpha|=k} C_\alpha \int\int\\ &\times \sum_{i,j,h=1,2} G_{i,j,h}(\xi^1,\eta,\zeta)g_i(\eta^1-\zeta^1-\xi^1,\eta^2-\zeta^2,\zeta,\xi^1)h_j(\xi^1,\zeta)v_h(\zeta)d\xi^1 d\zeta.\end{aligned}$$

We only need to show, from Proposition 2.2, that $G_{i,j,h} \in L^2(d\xi^1 d\zeta)$ uniformly in η. The fact that this is indeed so follows from the hypotheses and an analysis of eight different pieces of the kernel. The arguments are similar to the proof of Theorem 2.2 or the proof of the local part. For simplicity, we leave it to the interested reader. □

A simple partition of the unity argument and an application of Leibnitz's rule as in [8] (the proof of Corollary 1.6) give the following result.

COROLLARY 2.1. *Let $n_0/2 < s$, $0 \le l \le s$, $1 \le n_0 \le n$ $s \le s_b$, $n/2 < s$, and $q < \min\{s+l-n_0/2, s_b+s-n/2\}$. Let $k \ge 0$ be an integer, and for $|\alpha| \le k$*

$$a(x^1,\xi) \in S^{m+k;s+k,q+k}_{n_0}(\gamma), \qquad \partial^\alpha_\xi b(x,\xi) \in S^{-|\alpha|;s_b,q}(\Gamma).$$

Then $b \circ a(x,\xi)$, the symbol of operator $b(x,D)a(x^1,D)$, satisfying

$$\left\{ b\circ a(x,\xi) - \sum_{|\alpha|<k} \frac{1}{\alpha!}\partial^\alpha_\xi b D^\alpha_{x^1} a(x^1,\xi) \right\} \in S^{m;s,q,l,q}_{bd}(\Gamma).$$

The proof of Lemma 2.1 may be sharpened in a way similar to the proof of Lemma 1.12 in [8] to yield the following representation.

LEMMA 2.2. *Let $n_0/2 < s$, $0 \le l \le s$, $1 \le n_0 \le n, s$, and $q \le s+l-n_0/2$. Assume also that $k \ge 0$ is an integer and $\mu \ge 0$,*

$$a(x^1,\xi) \in S^{m;s+k,q+k}_{n_0}(\gamma), \qquad \partial^\alpha_\xi b(x,\xi) \in S^{k+\mu}\,.$$

Then there exist symbols $r(x,\xi)$, $\tilde{r}(x,\xi) \in S^{m;s,q,l,q}_{bd}(\Gamma)$ and $p(x,\xi), \tilde{p}(x,\xi) \in S^\mu_{1,0}$, such that

$$\left\{ b\circ a(x,D) - \sum_{|\alpha|<k} \frac{1}{\alpha!}\partial^\alpha_\xi b D^\alpha_x a(x,D) \right\} = r(x,D)p(x,D) + \tilde{p}(x,D)\tilde{r}(x,D)\,.$$

Accordingly, one can also prove a consequence of Lemma 2.2.

COROLLARY 2.2. *The conclusions of Lemma 2.2 hold if the assumptions on a and b are replaced with*

$$a(x^1,\xi) \in S^{m+k;s+k,q+k}_{n_0}(\gamma), \qquad \partial^\alpha_\xi b(x,\xi) \in S^\mu.$$

LEMMA 2.3. *Let $k \geq 1$, $n_0 < s$, $0 \leq l \leq s, q < s + l - n_0/2$, and*

$$a(x^1,\xi) \in S^{k;s+k,q+k}_{n_0}(\gamma).$$

Then the adjoint of a is an operator with symbol

$$a^*(x,\xi) \in S^{k;s+k,r+k}_{n_0}(\gamma)$$

and

$$\{a^*(x,\xi) - \bar{a}(x,\xi)\} \in S^{k-1;s+k-1,q+k-1,l,q}_{bd}(\gamma)\ .$$

The proof follows from the calculus of smooth $\psi.d.o.$ and Theorem 2.2 by observing that

$$a = \sum a_l(x^1)P_l(x^1,\xi)\ , \qquad P_{k,l} \in S^k$$

implies that

$$a^*(x,D) = \sum P^*_l(x,D)\bar{a}_l(x)$$

with $P^*_l(x,\xi) \in S^k$.

Finally, to complete the proof of Theorem 2.1, we need a Gårding type inequality. Let $\Pi : (x^1,x^2,\xi) \to (x^1,\xi)$ be a projection map.

LEMMA 2.4. *Let $n_0/2 < s$, $0 < \epsilon \leq 1$, $x = (x^1,x^2) \in \mathbf{R}^{n_0} \times \mathbf{R}^{n-n_0}$, and*

$$P(x,\xi) = a(x^1,\xi)P_0(x,\xi) + P_1(x,\xi) \geq 0, \tag{2.2}$$

where $a(x^1,\xi) \in S^{\epsilon,s+2\epsilon}_{n_0}(V)$, $P_0(x,\xi) \in S^0(U)$, and $P_1(x,\xi) \in S^\epsilon(U)$ for $V = \Pi U$. Then for all $u \in H^0_{\text{comp}}(U)$

$$Re\langle P(x,D)u,u\rangle \geq -C||u||^2_{H^0(U)}\ .$$

Proof. The proof may be given by following the general outline of the proof of Lemma 3.1 in [8] with some necessary modifications. As in Taylor [24], let $b(D,x,D)$ be the Friedrichs symmetrization of P. Set $R(x,D) = b(D,x,D) - p(x,D)$. If one can show that

$$R(x,D) : H^0_{\text{comp}} \to H^0_{\text{loc}}$$

is a bounded operator, then (2.2) would follow since $\langle b(D,x,D)u,u\rangle \geq 0$.

Let q be a smooth nonnegative even function, supported in $|\xi| \leq 1$ satisfying $\int q^2(\xi)d\xi = 1$, and define

$$F(\xi,\zeta) = \frac{1}{\langle\xi\rangle^{n/4}} q\left(\frac{\zeta-\xi}{\langle\xi\rangle^{1/2}}\right),$$

$$b(\eta,x,\xi) = \int F(\eta,\zeta)P(x,\zeta)F(\xi,\zeta)d\zeta.$$

Then

$$(b(D,x,D)u)\hat{}(\eta) = \int \hat{b}(\eta, \eta-\xi, \xi)\hat{u}(\xi)d\xi\ .$$

Thus

$$(R(x,D)u)\hat{}(\eta) = \int \hat{r}(\eta-\xi,\xi)\hat{u}(\xi)d\xi,$$

where $\hat{R}(\eta,\xi) = \hat{b}(\eta+\xi,\eta,\xi) - \hat{P}(\eta,\xi)$.

If one can show that

$$\langle\eta^1\rangle^s\langle\eta^2\rangle^r|\hat{R}(\eta,\xi)| \le Cg(\eta,\xi) \tag{2.3}$$

with $g \in L^2(d\eta)$ uniformly in ξ for some $r > (n-n_0)/2$, then similar arguments as in the proof of Proposition 2.1 or Proposition 2.2 yield

$$Ru \in H^0.$$

Observe that since $\int F^2(\xi,\zeta)d\zeta = 1$, we have

$$\hat{R} = \int F(\eta+\xi,\zeta)\hat{P}(\eta,\zeta)F(\xi,\zeta)d\zeta - \hat{P}(\eta,\xi) \tag{2.4}$$

$$= \int F(\eta+\xi,\zeta)\{\hat{P}(\eta,\zeta) - \hat{P}(\eta,\xi)\}F(\xi,\zeta)d\zeta$$

$$+ \int \{F(\eta+\xi,\zeta) - F(\xi,\zeta)\}\hat{P}(\eta,\xi)F(\xi,\zeta)d\xi\ . \tag{2.5}$$

On support of $F(\xi,\zeta)$, $|\zeta-\xi| \le \langle\xi\rangle^{1/2}$, hence for large ξ one has $\zeta \approx \xi$. Then (2.4) and the hypotheses on a, P_0, and P_1 imply that

$$|\hat{R}(\eta,\xi)| \le \frac{\langle\xi\rangle^\epsilon}{\langle\eta^1\rangle^{s+2\epsilon}\langle\eta^2\rangle^r} g_0(\eta,\xi)$$

for any r with $g_0(\eta,\xi) \in L^2(d\eta)$ uniformly in ξ.

The estimate (2.3) can then be established by following the proof of Lemma 3.1 in [8]. □

3. Proof of the main theorem. We now prove the main result of this paper, Theorem 1.1, a trace regularity theorem for the linear acoustic wave equation with nonsmooth coefficients in the principal part.

As we mentioned earlier, since the hypersurface $\{x_n = 0\}$ is a time-like surface, the method of energy estimates cannot be applied directly. Following [3], we alter the wave operator $\Box$ by a microlocal cutoff technique so that $\{x_n = 0\}$ will become a space-like surface with respect to the new operator. More precisely, we shall construct a $\psi.d.o.$ equation which is strictly hyperbolic corresponding to the trace $\{x_n = 0\}$. Since the operator in our construction is differential in x_n, the standard method of energy estimates (for example in [14] or [24]) can then be employed to derive the basic estimate. Naturally, there will be a remainder term. We then show that the remainder term can be controlled by the microlocal hypotheses and Theorem 2.1.

Let $\tilde{\Pi}_2 : X \subset T^*(\mathbf{R}^k) \to Y \subset \mathbf{R}^k \times \mathbf{R}^{k_0}$ serve as a map for $k > k_0$,

$$\tilde{\Pi}_2(X) = \{(x,y,\xi) \in Y : (x,y,\xi,\eta) \in X\}\ .$$

Proof of Theorem 1.1. Let γ_0, γ_1 be two conic subsets of the set $\{(x,t,\xi',\omega) \in \mathbf{R}^{n+1} \times \mathbf{R}^n; (x,t) \in \Omega_0,\ (\xi',\omega) \in \mathbf{R}^n, d(x)|\omega| \geq |\xi'|\}$ where Ω_0 is a small compact neighborhood of Ω, and $\gamma_0^{cl} \subset \gamma_1^{int}$. Then, construct an operator $Q \in C^\infty(\mathbf{R}, OPS^0(\mathbf{R}^n))$, $q = q(x,t,\xi',\omega)$, whose symbol satisfies

- $ES(Q) \subset \gamma_1$ and $0 \leq q_0 \leq 1$,
- $q_0 = 1$ on $\gamma_0 \cap \{(x,t,\xi',\omega),\ |(\xi',\omega)| > 1\}$,

where $q_0(\xi',\omega)$ is the principal symbol of Q. Define another operator E as

$$E = Q\Box^d_{x',t} + (I-Q)\Delta^d_{x',t}\ ,$$

where $\Box^d_{x',t} = d^2\partial_t^2 - \partial^2_{x'}$ and $\Delta^d_{x',t} = d^2\partial_t^2 + \partial^2_{x'}$.

Observe that the principal symbol of E

$$e_0 = q_0(d^2\omega^2 - |\xi'|^2) + (1-q_0)(d^2\omega^2 + |\xi'|^2) \geq C(\omega^2 + |\xi'|^2)\ ,$$

for $|(\omega,\xi')| \geq \delta$, with some positive constants C, δ. Hence, E is an elliptic $\psi.d.o.$ of order *two*.

Let $\phi = \phi(x,t) \in C_0^\infty(\mathbf{R}^{n+1})$ with supp $\phi \subseteq \{|x_n| \leq \epsilon_0\}$. We then have a strictly symmetric hyperbolic problem:

$$\begin{aligned}(-\partial^2_{x_n} + E)\phi u &= \Box_d \phi u + (I-Q)(\Delta^d_{x',t} - \Box^d_{x',t})\phi u \\ &= [\Box_d, \phi]u + \phi f + \phi\nabla\sigma\cdot\nabla u + 2(I-Q)\partial^2_{x'}\phi u \\ &= d^2[\partial_t^2, \phi]u - [\Delta, \phi]u + \phi f + \phi\nabla\sigma\cdot\nabla u + 2(I-Q)\partial^2_{x'}\phi u,\end{aligned} \tag{3.1}$$

where $\Box_d = d^2\partial_t^2 - \Delta$.

Thus, we obtain a standard wave equation with x_n playing the role of "time." This together with the fact that ϕ is compactly supported gives us a symmetric hyperbolic Cauchy problem with zero Cauchy data. It follows from a hyperbolic energy estimate in Taylor [24, pp. 73-78] that

(3.2)

$$\begin{aligned}\|(\phi u)\,|_{x_n=0}\,\|_l &\leq C\|\text{ r.h.s. of (3.1)}\|_{l-1} \\ &\leq C[\|\phi_1 u\|_l + \|\phi f\|_{l-1} + \|\phi\nabla\sigma\cdot\nabla u\|_{l-1} + \|(I-Q)\partial^2_{x'}\phi u\|_{l-1,\Omega_1}],\end{aligned}$$

where $\phi_1 \in C_0^\infty$, and $\phi_1 > 0$ on supp(ϕ), and supp$(\phi) \subset\subset \Omega_1$. Let $\tilde\phi \in C_0^\infty(\Omega_1)$ and $\tilde\phi = 1$ on Ω_2, where supp$(\phi) \subset\subset \Omega_2 \subset\subset \Omega_1$. Then the above estimate leads to

$$\|(\phi u)\,|_{x_n=0}\,\|_l \leq C[\|\phi_1 u\|_l + \|\phi f\|_{l-1} + \|\phi\nabla\sigma\cdot\nabla u\|_{l-1} + \|\tilde\phi(I-Q)\partial^2_{x'}\phi u\|_{l-1,\Omega_1}]$$

where we have used $||(1-\tilde\phi)(I-Q)\partial^2_{x'}\phi u||_{l-1,\Omega_1} \leq C||\phi u||_r$ for any real r.

Using the hypotheses, similar to the proposition in Beals and Reed [7], one can show that $u \in H^l_{\text{loc}}$. Hence a generalized Schauder's lemma (Lemma 2.2 in [3]) yields

$$\phi\nabla\sigma\cdot\nabla u \in H^{l-1}\ .$$

Therefore, to complete the proof it suffices to show that

$$(I-Q)\partial^2_{x'}\phi u \in H^{l-1}_{\text{loc}}(\Omega_1)\ ,$$

which requires the use of Lemma 3.1 stated below. In order to apply Lemma 3.1, we choose $B = I - Q \in C^\infty(\mathbf{R}^1, OPS^0(\mathbf{R}^n))$ of order $m = 0$, $A = \Box_d$ of order $m_0 = 2$, and $h = l - 2$ in the statement of Lemma 3.1.

Concerning the assumption (1) of Lemma 3.1, $Ell(A)$ defined as the microlocal elliptic region of $A = \Box_d$ is easy to determine by knowing that the operator $\Box_d$ is elliptic away from the set $\{d^2\omega^2 = |\xi|^2\}$. To verify hypothesis (3), one only needs to look at

$$\Box_d \partial_{x'}^2 \phi u = [\Box_d, \partial_{x'}^2 \phi] u + \partial_{x'}^2 \phi (\nabla\sigma \cdot \nabla u + f) \, .$$

Then $\Box_d \partial_{x'}^2 \phi u \in H^{l-3}$ follows by some simple commutator arguments. Therefore, the only assumption that needs to be checked is

$$u \in H^{l+1}_{m\ell}([T^*(\mathbf{R}^{n+1}) \backslash Ell(\Box_d)] \cap \tilde{\Pi}_2^{-1} ES(I-Q))$$

and this demands Theorem 2.1.

In the statement of Theorem 2.1 choose

$$(m, n_0, n, l, s, q, \epsilon) = (0, n, n+1, l-1, s-2, l, 1)$$

then the microlocal hypotheses verify all the assumptions of Theorem 2.1. Notice that the main assumption, $s > 3 + n/2$, is required by the corresponding hypothesis ("$q < l + s - n_0/2$) in the statement of Theorem 2.1. Let γ_0 and γ_1 approach the set $\{(x,t,\xi,\omega) : \ (x,t) \in \Omega_0 \ , \ \ d|\omega| \geq |\xi'|\}$. The set

$$[T^*(\mathbf{R}^{n+1}) \backslash Ell(\Box)] \cap \tilde{\Pi}_2^{-1} ES(I-Q)$$

is contained in a small (conic) neighborhood of the Hamiltonian flow out of γ. Hence Theorem 2.1 and the microlocal initial hypotheses yield that

$$u \in H^{l+1}_{m\ell}([T^*(\mathbf{R}^{n+1}) \backslash Ell(\Box)] \cap \tilde{\Pi}_2^{-1} ES(I-Q)) \, .$$

We have completed the proof of Theorem 1.1. □

For the sake of completeness, we state a lemma given in [2].

Consider a smooth family of $\psi.d.o.$ $P \in C^\infty(\mathbf{R}^{k-k_0}, OPS^m(\mathbf{R}^{k_0}))$, i.e., for each $y \in \mathbf{R}^{k-k_0}$ with $k_0 < k$, $P(x,y,D_x) \in OPS^m(\mathbf{R}^{k_0})$. It is known, see [2] or [23], that P is not necessarily a $\psi.d.o.$ in $\mathbf{R}^k$. However, as shown in [2], a smooth family of $\psi.d.o.$ behaves like a $\psi.d.o.$

Recall the map, for $k > k_0$,

$$\tilde{\Pi}_2(X) = \{(x,y,\xi) \in Y \subset \mathbf{R}^k \times \mathbf{R}^{k_0} : \ (x,y,\xi,\eta) \in X \subset T^*(\mathbf{R}^k)\} \, .$$

The normal bundle of a foliation $\mathbf{R}^k = \mathbf{R}^{k-k_0} \times \mathbf{R}^{k_0}$ is the set

$$\mathcal{N} = \{(x,y,\xi,\eta) \in \mathbf{R}^{k_0} \times \mathbf{R}^{k-k_0} \times \mathbf{R}^{k_0} \times \mathbf{R}^{k-k_0}, \xi = 0\} \, .$$

LEMMA 3.1 (Fattening lemma). *Let $B(x,y,D_x) \in C^\infty(\mathbf{R}^{k-k_0}, OPS^m(\mathbf{R}^{k_0})$ and $A(x,y,D_x,D_y) \in OPS^{m_0}(\mathbf{R}^k)$, where $1 \leq k_0 \leq k$. Let*

$$\mathcal{N} = \{(x,\xi) \in \mathbf{R}^k \times \mathbf{R}^k, (\xi_1, \cdots, \xi_{k_0}) = 0\}$$

be the normal bundle of $\mathbf{R}^{k_0} \times \mathbf{R}^{k-k_0}$. Also, assume that

(1) *A is microlocal elliptic on a conic set $Ell(A)$, with $\mathcal{N} \subset\subset Ell(A)$;*
(2) *$u \in H^h \cap H^{h+1}_{m\ell}([T^*(\mathbf{R}^k) \backslash Ell(A)] \cap \tilde{\Pi}_2^{-1} ES(B(\cdot,y,\cdot)))$;*
(3) *$A\phi u \in H^{h-m_0+1}_{\text{loc}}(\mathbf{R}^k)$, where $\phi(x) \in C_0^\infty(\mathbf{R}^k)$.*

Then

$$B\phi u \in H^{h-m+1}_{\text{loc}}(\mathbf{R}^k) ,$$

in addition, if B is either a convolutional operator or its symbol has compact support in spatial variables,

$$B\phi u \in H^{h-m+1}(\mathbf{R}^k) .$$

REFERENCES

[1] G. Bao, *Time like trace regularity of wave equations*, in Second International Conference on Mathematical and Numerical Aspects of Wave Propagation, R. Kleinman et al., eds., Society for Industrial and Applied Mathematics, Philadelphia, 1993, pp. 20–29.

[2] G. Bao and W. Symes, *A trace theorem for solutions of linear partial differential equations*, Math. Meth. Appl. Sci., 14 (1991), pp. 553–562.

[3] ———, *Trace regularity for a second order hyperbolic equation with nonsmooth coefficients*, J. Math. Anal. Appl., 174 (1993), pp. 370-389.

[4] M. Beals, *Self-spreading and strength of singularities for solutions to semilinear wave equation*, Ann. of Math., 118 (1983), pp. 187–214.

[5] ———, *Propagation of smoothness for nonlinear second order strictly hyperbolic equations*, Proc. Pure Math., 43 (1985), pp. 21–44.

[6] ———, *Propagation and Interaction of Singularities in Nonlinear Hyperbolic Problems*, Birkhäuser-Verlag, Boston, MA, 1989.

[7] M. Beals and M. Reed, *Propagation of singularities for hyperbolic pseudodifferential operators and applications to nonlinear problems*, Comm. Pure Appl. Math., 35 (1982), pp. 169–184.

[8] ———, *Microlocal regularity theorems for nonsmooth pseudodifferential operators and applications to nonlinear problems*, Trans. Amer.Math. Soc., 285 (1984), pp. 159–184.

[9] J. -M. Bony, *Calcul symbolique et propagation des singularitéaires pour les équations aux dérivées partielles non-linéaires*, Ann. Sci. École Norm. Sup., 14 (1981), pp. 209–246.

[10] J. Y. Chemin, *Interaction contrôlée dans les équations aux dérivées partielles non linéaires*, C. R. Acad. Sci. Paris Sér. Math., 303 (1986), pp. 451–453.

[11] L. Hörmander, *On the existence and regularity of solutions of linear pseudo-differential equations*, Enseign. Math., 17 (1971), pp. 121–133

[12] ———, *Linear differential operators*, Actes Congr. Internat. Math., Vol.1 (Nice, 1970); Gauthier-Villars, Paris, 1971, pp. 121–133.

[13] J. U. Kim, *Trace regularity in the boundary control of a wave equation*, preprint, 1992.

[14] F. John, *Partial Differential Equations*, Springer-Verlag, New York, 1982.

[15] P. Joly, *Some trace theorems in anisotropic Sobolev spaces*, SIAM J. Math. Anal., 23 (1992), pp. 799–819.

[16] I. Lasiecka and R. Triggiani, *Trace regularity of the solutions of the wave equation with homogeneous Neumann boundary conditions and compactly supported data*, J. Math. Anal. Appl., 14 (1989), pp. 49–71.

[17] ———, *Sharp regularity theory for second order hyperbolic equations of Neumann type*, Anna. Mat. Pure et Appl. (IV) Vol. CLVII (1990).

[18] ———, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometric conditions*, Appl. Math. Optim., 25 (1992), pp 189–224.

[19] L. Nirenberg, *Lectures on Linear Partial Differential Equations*, CBMS Regional Conf. Ser. in Math., No. 17, Amer. Math. Soc., Providence, RI, 1973.

[20] J. Rauch, *Singularities of solutions to semilinear wave equations*, J. Math. Pures Appl., 58 (1979), pp. 299–308.

[21] J. Rauch and M. Reed, *Nonlinear microlocal analysis of semilinear hyperbolic systems in one space dimension*, Duke Math. J., 49 (1982), pp. 379–475.

[22] W. Symes, *A trace theorem for solutions of the wave equation, and the remote determination of acoustic sources*, Math. Meth. Appl. Sci., 5 (1983), pp. 131–152.

[23] M. Taylor, *Reflection of singularities of solutions to systems of differential equations*, Comm. Pure Appl. Math., 28 (1975), pp. 457–475.

[24] ———, *Pseudo-Differential Operators*, Princeton Univ. Press, Princeton, NJ, 1981.

[25] ———, *Pseudodifferential Operators and Nonlinear PDE*, Birkhäuser-Verlag, Boston, MA, 1991.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 147–179, January 1995

008

UNIFORM STRICT CONVEXITY OF A COST FUNCTIONAL FOR THREE-DIMENSIONAL INVERSE SCATTERING PROBLEM *

MICHAEL V. KLIBANOV† AND OLGA V. IOUSSOUPOVA‡

Abstract. An inverse problem of determination of the coefficient $a(x)$ in the equation $u_{tt} = \Delta u + a(x)u, x \in \mathbb{R}^3, t \in (0,T)$ is considered with initial conditions $u(x,0) = 0, u_t(x,0) = \delta(x)$, and some additional data that can be treated as *backscattering* information. The goal is to develop a finite-dimensional technique that would be a basis for future computations. We reduce our inverse scattering problem to an equivalent Cauchy problem for a nonlinear hyperbolic integrodifferential equation with the data on the lateral side of a time cylinder. It is assumed that the solution $v(x,t)$ of this equation has the form $v(x,t) = p(x,t) + w(x,t)$, where function $p(x,t)$ is given and function $w(x,t)$ is unknown and has a finite number of nonzero Fourier coefficients. In particular, function $p(x,t)$ can be considered as a first guess. A special cost-functional $J_\lambda(w)$ dependent on a large parameter λ is introduced. The main result of this paper is Theorem 1.1. By this theorem, the functional J_λ is *uniformly strictly convex* on any ball B with the center at the origin with a proper choice of the parameter $\lambda = \lambda(B)$. Therefore, by this theorem, a finite-dimensional perturbation of a true solution of the above-mentioned nonlinear Cauchy problem can be found by convex minimization techniques.

Key words. inverse scattering problem, Carleman estimates, strict convexity, global convergence, imaging, strong scattering

AMS subject classification. 35R30

0. Introduction.

Numerical methods for three-dimensional inverse scattering problems (ISP) currently attract considerable interest. In part, it is due to their significance for imaging of the *complicated internal* structure of inhomogeneous media, such as biological tissues, industrial devices, etc. One of the most attractive advantages of such imaging is in displaying the subtle structures of abnormal inclusions of the media, which is of special importance for diagnosis of some human diseases, noninvasive quality control of industrial devices, etc. Presumably an imaging device should introduce electromagnetic waves into inhomogeneous media. Then this device should measure scattering data all around the media and treat this data computationally. Finally, it should display the image of the internal structure of the media on the basis of computational results. Thus numerical methods for three-dimensional ISPs play a fundamental role in imaging processes. We also refer to such traditional areas of application of three-dimensional ISPs as ocean acoustics and geophysics.

Currently, these methods have been developed mainly in a frequency domain, i.e., for the Helmholtz-like equation $\Delta u + k^2(1 + c(x))u = 0$; cf. Cheney [3], Colton and Monk [4], [5] and Gutman and Klibanov [6]; [3] and [4] contain surveys. However, ISPs with time-dependent data play a substantial role in imaging problems as well. Numerical methods for three-dimensional ISP with time-dependent data have been studied by Baylis, Li, and Morawetz [1] and by Klibanov and Malinsky [13]. ISPs for the equation $u_{tt} = \Delta u + a(x)u, x \in \mathbb{R}^3$ were considered in [1] and [13]. Here

* Received by the editors February 8, 1993; accepted for publication (in revised form) September 21, 1993.

† Department of Mathematics, University of North Carolina at Charlotte, Charlotte, North Carolina 28223. The research of this author was supported by Office of Naval Research grant N00014-92-J-1008.

‡ Department of Mathematics, Samara Pedagogical Institute, Samara, 443099, Russia.

potential $a(x)$ is the unknown function. Note that this equation can be obtained by the Fourier transform from the Schrödinger equation $\Delta V + k^2 V + a(x)V = \delta(x)$ which is commonly used to model many-body systems. A Newton–Kantorovich numerical method was developed in [13] for the weak scattering case, which is in the case when $|a(x)| \ll 1$.

Newton–Kantorovich methods, however, require that a distance between the starting point (i.e., the initial guess) and the solution has to be sufficiently small, which is, of course, not always the case in applications. We note that in practical computations of inverse problems the solutions are usually sought as m-dimensional perturbations of a certain initial guess. Thus it would be very helpful to develop a finite-dimensional technique which would have a *guaranteed* convergence for any fixed dimension m and for any distance between the initial guess and the solution of the ISP. Such a technique would be a basis for future practical computations which would give results sufficient for practical goals. One should expect, however, that the question of convergence of such a method as $m \to \infty$ would be very hard to answer due to *ill-posedness* of the inverse problems; cf. Lavrent'ev et al. [18] and Tikhonov and Arsenin [23] (see also §6).

In this paper we develop such a finite-dimensional technique for an ISP for the above mentioned hyperbolic equation. Our ISP deals with the *backscattering* data. First we get rid of the unknown coefficient $a(x)$ and obtain, in this way, a time-like Cauchy problem (1.21)–(1.23) for a nonlinear hyperbolic integrodifferential equation. This problem is actually equivalent with the original ISP. Let $N \geq 1$ be an arbitrary integer and $m = N^3$. We seek a solution $v(x,t)$ of this nonlinear Cauchy problem in the form $v(x,t) = w(x,t) + p(x,t)$, where function $p(x,t)$ is given, function $w(x,t)$ is unknown, and $w(x,t)$ has m nonzero Fourier coefficients with respect to a certain orthonormal basis. Therefore function $p(x,t)$ can be considered as a first guess for the solution $v(x,t)$. But we do not impose "smallness" conditions on the function $w(x,t)$.

Let function $w_*(x,t)$ be the sought solution and $c_* = c(w_*) \in \mathbb{R}^m$ be the vector of the Fourier coefficients of the function w_*. By Tikhonov's principle one can assume that a solution of an ill-posed problem belongs to an a priori given compact set; see [23]. Thus without loss of generality, we can assume that $c_* \in B_{1/2}$ where

$$B_{1/2} = \{y \in \mathbb{R}^m : |y| < \tfrac{1}{2}\}, \qquad B_1 = \{y \in \mathbb{R}^m : |y| < 1\}.$$

Then we introduce a special weighted minimizing cost functional $J_\lambda = J_\lambda(w)$, where λ is a large positive parameter. The functional J_λ is tightly connected with *Carleman estimates*; cf. Klibanov [12]. Vector $c_* = c(w_*)$ provides a global minimum of J_λ on $\bar{B}_1$, in the case of the absence of a noise in the data (regarding the noise, see Theorem 1.3).

The main result of this paper is Theorem 1.1, which claims that there exists a constant $\tilde{\lambda} = \tilde{\lambda}(B_1)$ such that the functional J_λ is *uniformly strictly convex* on $\bar{B}_1$. The *genuine* meaning of this result is that one can find a finite-dimensional perturbation of the true solution of the nonlinear Cauchy problem (1.21)–(1.23) by convex minimization methods. We note that Theorem 1.1 guarantees a *global* convergence of a number of traditional minimization algorithms to the global minimum w_* of J_λ (on $\bar{B}_1$). In particular, we prove, basically for the sake of completeness, Theorem 1.2 which establishes a convergence rate of the simplest version of the gradient method, in a case when the starting point is the center of B_1 and the data does not contain a noise. We also prove Theorem 1.3, which shows that the gradient method provides a "reasonable" solution even in the presence of a noise in the data. Finally, Theorem

1.1 implies Theorem 1.4, which claims that our "m-dimensional" solution $w(x,t)$ is unique. Of course, Theorem 1.4 cannot be considered as a global uniqueness result (see a discussion of the uniqueness issue below). One can argue, however, that sometimes computations are successfully carried out despite the lack of global uniqueness results. The reason for this is that theorems similar to Theorem 1.4 are sometimes sufficient for practical computations and provide good indications of the validity of global uniqueness results. We consider this paper as a basis for future computations, which will be discussed elsewhere (see also §6).

To our knowledge, Symes was the first one who proved, using an entirely different technique, uniform convexity of a minimizing cost functional for an inverse problem for the equation $u_{tt} = c^2(x_3)\Delta u + \delta(x)f(t)$ with unknown coefficient $c(x_3)$; see, e.g., Theorem 6.1 in [22] and the references cited there. He also raised a question about similar results in a case when the unknown coefficient depends on $n = 3$ variables. As far as we know our result is the first one in this direction.

Let T = constant > 0 and $\mathbb{R}^3_T = R^3 \times (0,T)$. Let a nonnegative function $a \in C^2(\mathbb{R}^3)$ and a function $u(x,t)$ be a solution of the Cauchy problem

$$\begin{aligned} &u_{tt} = \Delta u + a(x)u \quad \text{in } \mathbb{R}^3_T, \\ &u|_{t=0} = 0, \qquad u_t|_{t=0} = \delta(x), \end{aligned} \tag{0.1}$$

where $\delta(x)$ is the delta function. Introduce spherical coordinates

$$x_1 = r\cos\theta\sin\varphi, \quad x_2 = r\sin\theta\sin\varphi, \quad x_3 = r\cos\varphi, \theta \in [0,2\pi), \qquad \varphi \in [0,\pi).$$

Let r_0, R be positive constants such that $r_0 < R < \frac{1}{3}T$. Let $\Omega_{r_0} \subset \mathbb{R}^3$ be a cylinder,

$$\Omega_{r_0} = \left\{ x \in \mathbb{R}^3 \colon \sqrt{x_1^2 + x_2^2} = r\sin\varphi < r_0 \right\}.$$

Let ω be the boundary of Ω_{r_0}.

The principal problem of interest for us is as follows: *Assume that the function $a(x) = 0$ for $x \in \Omega_{r_0}$ and is unknown otherwise. We would like to determine this function for $x \in (\mathbb{R}^3|\Omega_{r_0}) \cap \{/x/ < R\}$ assuming that the following function ψ is given*:

$$u/_\omega = \psi(x,t), \qquad t \in (0,T). \tag{0.2}$$

We note that since function $a(x) = 0$ inside of Ω_{r_0}, then we can uniquely determine function $u(x,t)$ for $(x,t) \in \Omega_{r_0} \times (0,T)$ as a solution of the following boundary value problem:

$$\begin{aligned} &u_{tt} = \Delta u \quad \text{in } \Omega_{r_0} \times (0,T), \\ &u|_{t=0} = 0, \qquad u_t|_{t=0} = \delta(x), \\ &u|_\omega = \psi(x,t). \end{aligned} \tag{0.3}$$

Hence actually the following function $\mu(x,t)$ is given:

$$u = \mu(x,t) \quad \text{for } (x,t) \in \Omega_{r_0} \times (0,T). \tag{0.4}$$

Remarks. (i) In fact, it would be more natural, perhaps, to assume that function $a(x) = 0$ only in the ball $x \in \{|x| < r_0\}$ and function $u(x,t)$ is given only on the

sphere $|x| = r_0$. In this case, however, we would face some specific difficulties related to the fact that the Jacobian

$$\det\left(\frac{\partial(x_1, x_2, x_3)}{\partial(r, \theta, \varphi)}\right) = r^2 \sin\varphi$$

equals zero at $\varphi = 0, \pi$, i.e., on the line $x_1 = x_2 = 0$ (see also the second remark in §3). We would face these difficulties only because we get Carleman estimates in terms of spherical coordinates (not cartesian ones). Consequently, we think that a modification of our Carleman estimates would lead to an elimination of these obstacles. Currently, however, we do not know how to handle this.

(ii) Nevertheless condition (0.2) is not a very restrictive one. For instance, this condition is valid in the case when the function $a(x)$ has a compact support in a neighborhood of a point $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3)$ such that $\sqrt{\bar{x}_1^2 + \bar{x}_2^2} > 2r_0$. That is, $a(x) = 0$ for $|x - \bar{x}| > \sigma$, where $0 < \sigma < |\bar{x}| - r_0$. In this case one can assume, for instance, that one measures scattering data on a sphere $|x - \bar{x}| = \sigma$ which certainly can be the case in imaging problems. Then one can get function ψ in (0.2) by solving a boundary value problem similar with (0.3).

(iii) In fact, measurements (0.2) on the surface ω contain backscattering information, which is the only unique "useful" part of the measuring signal. Indeed, waves propagate from the source $x = 0$, which is placed inside of the cylinder Ω_{r_0}, and support of the unknown function $a(x)$ is outside of Ω_{r_0}. If our measurement data did not contain the backscattering signal, we would be unable to get a uniqueness result (see Theorem 1.4 below).

It is widely known that numerical methods for ill-posed problems, and for inverse problems in particular, are closely connected with uniqueness theorems for these problems. It is also well known that global uniqueness results are much more difficult to get in a case where unknown coefficients depend on $n \geq 2$ variables than those for a $n = 1$ case. In particular, uniqueness theorems for the inverse problem (0.1) and (0.2), as well as for closely connected problems of integral geometry, are currently proven only under rather restrictive conditions imposed on $a(x)$; see Lavrent'ev, Romanov, and Shishatskii [18] and Romanov [19]. To our knowledge, the question about the *global* uniqueness theorem for this problem still remains open despite a large number of attempts to answer it. Incidentally, the lack of methods of proofs of global uniqueness results is, perhaps, the second major reason (in addition to ill-posedness) why we cannot yet investigate our functional $J_\lambda(w)$ as $m \to \infty$.

Nevertheless, global uniqueness results were proved for similar multidimensional inverse problems under the assumption that $u(x, 0) \neq 0$ for all "needed" x. These results were proven by the method of Carleman estimates which has been applied to inverse problems beginning from the work of Bukhgeim and Klibanov in 1981 [2], see also the survey in Klibanov [12] and references cited therein. Recently Klibanov et al. started to apply Carleman estimates for proofs of convergence of numerical methods both for ill-posed Cauchy problems and for inverse problems; see [11] and [13]–[15]. The method of the current paper is also based on Carleman estimates. Note that by the classical approach, these estimates are applied to proofs of uniqueness and stability results for ill-posed Cauchy problems; cf. Hörmander [8], Isàkov [9], and Lavrent'ev et al. [18].

The paper is constructed as follows. In §1 we show how to get a time-like Cauchy problem for nonlinear hyperbolic integrodifferential equation, instead of the inverse problem (0.1) and (0.2). In this section, we also introduce the functional J_λ and formulate Theorems 1.1–1.4. In §2 we prove miscellaneous lemmas. In §3 we prove

a special form of the Carleman estimate (Theorem 3.1). Section 4 is devoted to the proof of Theorem 1.1. In §5 we prove Theorems 1.2 and 1.3. Section 6 is devoted to the discussion of the results obtained in this paper.

We close the Introduction with basic notation. In the sequel, all the functions are real valued. Let $T_1 = T - R$. Denote

$$\mathbb{R}^3_{T_1} = \mathbb{R}^3 \times (0, T_1),$$
$$D_{T_1} = \left\{(x,t)\colon r_0 < |x| < R, \sin\varphi > \frac{r_0}{R}, t \in [0,T_1)\right\}.$$

Hence for all $(x,t) \in \bar{D}_{T_1}$,

$$r\sin\varphi \geq \frac{r}{R}\cdot r_0 \geq \frac{r_0^2}{R}$$

and the conic surface

$$\partial_1(D_{T_1}) = \left\{\sin\varphi = \frac{r_0}{R}, r_0 < |x| < R, t \in (0,T_1)\right\} \tag{0.5}$$

is a part of the boundary $\partial(D_{T_1})$ of the domain D_{T_1}. Furthermore $\partial_1(D_{T_1}) \subset \Omega_{r_0}$, because for $(x,t) \in \partial_1(D_{T_1})$

$$\sqrt{x_1^2 + x_2^2} = r\sin\varphi = \frac{r}{R}\, r_0 < r_0.$$

Hence by (0.4) the following functions are given:

$$u|_{\partial_1(D_{T_1})}, \qquad \nabla u|_{\partial_1(D_{T_1})}. \tag{0.6}$$

Consider the part $\partial_2(D_{T_1}$ of the boundary $\partial(D_{T_1})$,

$$\partial_2(D_{T_1}) = \left\{|x| = r_0, \sin\varphi > \frac{r_0}{R}, t \in (0,T_1)\right\}. \tag{0.7}$$

Denote

$$\eta = \partial_1(D_{T_1}) \cup \partial_2(D_{T_1}). \tag{0.8}$$

Hence by (0.4) the following two functions are given:

$$u|_\eta = \psi_1, \qquad \left.\frac{\partial u}{\partial n}\right|_\eta = \psi_2, \tag{0.9}$$

where n is an outward normal vector on η.

Let $H^2(D_{T_1})$ be the Sobolev space and $H_0^2(D_{T_1})$ be a subspace of $H^2(D_{T_1})$ such that for all $f \in H_0^2(D_{T_1})$,

$$f|_\eta = \left.\frac{\partial f}{\partial n}\right|_\eta = 0.$$

More generally, let P be a subdomain of D_{T_1} such that $\partial P \cap \eta = \eta(P) \neq \emptyset$, where ∂P is the boundary of P. Then $H_0^2(P)$ denotes a subspace of $H^2(P)$ such that

$$f|_{\eta(P)} = \left.\frac{\partial f}{\partial n}\right|_{\eta(P)} = 0,$$

for all $f \in H_0^2(P)$. Likewise, we introduce a Sobolev space $H_0^2(r_0, R)$ as a space of functions $f(r)$ having a finite norm

$$\|f\|_{H_0^2(r_0,R)} = \left\{ \int_{r_0}^{R} (|f|^2 + |f'|^2 + |f''|^2) r^2 \, dr \right\}^{1/2}$$

and such that $f(r_0) = f'(r_0) = 0$.

1. Statement of the main results. Solution u of the Cauchy problem (0.1) with $a \in c^2(\mathbb{R}^3)$ has the form

$$u(x,t) = \frac{\delta(t-|x|)}{4\pi|x|} + \hat{u}(x,t), \tag{1.1}$$

where function $\hat{u}(x,t) = 0$ for $t < |x|$, and $\hat{u} \in C^3(t \geq |x|)$. Since $a(x) \geq 0$, using the Kirchgoff formula one can simply derive that

$$\hat{u} \geq 0. \tag{1.2}$$

Consider the function

$$u_1 = \int_0^t u(x,\tau)\, d\tau. \tag{1.3}$$

Then (0.1), (0.8), and (1.1)–(1.3) imply

$$u_{1tt} = \Delta u_1 + a(x) u_1 \quad \text{in } \mathbb{R}^3_T, \tag{1.4}$$

$$u_1(x,t) = 0 \text{ for } t < |x|, \tag{1.5}$$

$$u_1|_{t=|x|} = \frac{1}{4\pi|x|}, \tag{1.6}$$

$$u_1|_\eta = \hat{\psi}_1(x,t), \qquad \left.\frac{\partial u_1}{\partial n}\right|_\eta = \hat{\psi}_2(x,t), \tag{1.7}$$

$$u_1 > 0 \quad \text{for } t \geq |x|, \tag{1.8}$$

$$u_1 \in C^3(t \geq |x|), \tag{1.9}$$

where functions $\hat{\psi}_1, \hat{\psi}_2$ have the form

$$\hat{\psi}_1(x,t) = \int_0^t \psi_1(x,\tau)\, d\tau, \qquad \hat{\psi}_2(x,t) = \int_0^t \psi_2(x,\tau)\, d\tau.$$

Now we want to replace the characteristic conic surface $\{t = |x|\}$ with the plane $\{t = 0\}$. In order to do that we introduce a new function $\tilde{u}(x,t) = u_1(x, t + |x|)$. Hence (1.4)–(1.7) implies

$$\Delta\tilde{u} - 2\tilde{u}_{rt} + \frac{2}{r}\,\tilde{u}_t + a(x)\tilde{u} = 0 \quad \text{in } \mathbb{R}^3_{T_1}, \tag{1.10}$$

$$\tilde{u}|_{t=0} = \frac{1}{4\pi|x|}, \tag{1.11}$$

$$\tilde{u}|_\eta = \hat{\psi}_1(x, t + r), \qquad \left.\frac{\partial\tilde{u}}{\partial n}\right|_\eta = \hat{\psi}_2(x, t + r). \tag{1.12}$$

Likewise, (1.8) and (1.9) lead to

$$\tilde{u} > 0 \text{ in } \mathbb{R}^3_{T_1}, \tag{1.13}$$

$$\tilde{u} \in C^3(\bar{\mathbb{R}}^3_{T_1}). \tag{1.14}$$

Since by (1.13) the function $\tilde{u}$ is positive, we can consider the function $\tilde{v} = \ln\tilde{u}$. Hence $\tilde{u} = e^{\tilde{v}}$. By (1.14), $\tilde{v} \in C^3(\bar{\mathbb{R}}^3_{T_1})$. Besides, (1.10)–(1.12) lead to

$$\Delta\tilde{v} - 2\tilde{v}_{rt} + |\nabla\tilde{v}|^2 - 2\tilde{v}_r\tilde{v}_t + \frac{2}{r}\,\tilde{v}_t + a(x) = 0, \tag{1.15}$$

$$\tilde{v}|_{t=0} = \ln\frac{1}{4\pi r}, \tag{1.16}$$

$$\tilde{v}|_\eta = \ln[\hat{\psi}_1(x, t + r)], \tag{1.17}$$

$$\left.\frac{\partial\tilde{v}}{\partial n}\right|_\eta = \left(\frac{\hat{\psi}_2}{\hat{\psi}_1}\right)(x, t + r). \tag{1.18}$$

Now we can eliminate function $a(x)$ from (1.15) simply by differentiation with respect to t. The price for this is that we will not know initial data for the function.

$$v = \tilde{v}_t. \tag{1.19}$$

Nevertheless by (1.16),

$$\tilde{v}(x,t) = \int_0^t v(x,\tau)\,d\tau + \ln\left(\frac{1}{4\pi r}\right). \tag{1.20}$$

Denote

$$\psi_3 = \frac{\partial}{\partial t}\ln[\hat{\psi}_1(x, t + r)],$$

$$\psi_4 = \frac{\partial}{\partial t}\left[\left(\frac{\hat{\psi}_2}{\hat{\psi}_1}\right)(x, t + r)\right].$$

Thus (1.15)–(1.20) lead to

$$L(v) = 0, \qquad v \in C^2(\bar{\mathbb{R}}^3_{T_1}),$$
$$v|_\eta = \psi_3, \qquad \left.\frac{\partial v}{\partial n}\right|_\eta = \psi_4. \tag{1.21}$$

Here the hyperbolic nonlinear integrodifferential operator L has the form

$$L(v) = L_0 v + 2\nabla v \int_0^t \nabla v(x,\tau)\,d\tau - 2v_t \int_0^t v_r(x,\tau)\,d\tau$$
$$- 2v_r v + \frac{2}{r} v_t + 2\nabla v \nabla g - 2v_t g'(r), \tag{1.22}$$

where

$$L_0 v = \Delta v - 2v_{rt}$$

and

$$g(r) = \ln\left(\frac{1}{4\pi r}\right). \tag{1.23}$$

Now assume that the function $v(x,t)$ is found. Then using (1.20) we find the function $\tilde{v}(x,t)$. Let $\tilde{u} = e^{\tilde{v}}$. Then by (1.10) and (1.11) we obtain

$$a(x) = 4\pi|x|(\Delta\tilde{u} - 2\tilde{u}_{rt})(x,0). \tag{1.24}$$

Therefore we have proven the following lemma.

LEMMA 1.1. *Solving of the inverse problem* (0.1) *and* (0.2) *is equivalent to solving of the time-like Cauchy problem* (1.21)–(1.23).

Below we always assume that function $v(x,t)$ in (1.21)–(1.23) has the form

$$v(x,t) = w(x,t) + p(x,t), \tag{1.25}$$

where function $p(x,t)$ is *given,*

$$p \in H^2(D_{T_1}) \quad \text{and} \quad p|_\eta = \psi_3, \qquad \left.\frac{\partial p}{\partial n}\right|_\eta = \psi_4. \tag{1.26a}$$

Hence $w \in H_0^2(D_{T_1})$. One could consider the function $p(x,t)$ as a first guess for the function $v(x,t)$.

In the sequel, we will assume that the function $w(x,t)$ in (1.25) has a finite number of Fourier coefficients with respect to an orthonormal basis. We describe conditions imposed on this basis below. Let $\varphi_1 = \arcsin(r_0/R), \varphi_2 = \pi - \varphi$ and $L_{2,\varphi}$ be a weighted L_2 space of functions $f(\theta,\varphi), \theta \in (0,2\pi), \varphi \in (\varphi_1,\varphi_2)$ having a finite norm

$$\|f\|_{L_{2,\varphi}} = \int_0^{2\pi} d\theta \int_{\varphi_1}^{\varphi_2} |f(\theta,\varphi)|^2 \sin\varphi\,d\varphi.$$

Furthermore, let $\tilde{H}_0^2(D_{T_1})$ be a Sobolev space of functions having a finite norm

$$\|f\|_{\tilde{H}_0^2(D_{T_1})} = \left\{ \int_0^{2\pi} d\theta \int_{\varphi_1}^{\varphi_2} \sin\varphi\,d\varphi \int_{r_0}^{R} r^2\,dr \right.$$
$$\left. \times \int_0^{T_1} dt[|f|^2 + |f_r|^2 + |f_t|^2 + |f_{rr}|^2 + |f_{tt}|^2] \right\}^{1/2}$$

and such that $f|_{r=r_0} = f_r|_{r=r_0} = 0$.

Choose an orthonormal basis $\{\phi_n(\theta,\varphi)\}_{n=1}^{\infty}$ in $L_{2,\varphi}$ such that the following conditions are fulfilled

$$\phi_n \in c^2([0,2\pi]\times[\varphi_1,\varphi_2]),$$

$$\text{(1.26b)} \qquad \phi_n(0,\varphi) = \phi_n(2\pi,\varphi), \qquad \frac{\partial\phi_n}{\partial\theta}(0,\varphi) = \frac{\partial\phi_n}{\partial\theta}(2\pi,\varphi),$$

$$\phi_n(\theta,\varphi_1) = \frac{\partial\phi_n}{\partial\varphi}(\theta,\varphi_1) = \phi_n(\theta,\varphi_2) = \frac{\partial\phi_n}{\partial\theta}(\theta,\varphi_2) = 0.$$

Only the third set of conditions (1.26b) seems a little bit strange among these because we deal with the L_2 space. For this reason let us briefly describe one *possible* example of an orthonormal basis in $L_{2,\varphi}$, which satisfies (1.26b). Let $\tilde{\phi}_n(\theta,\varphi)$ be an orthonormal basis in $L_2(0,2\pi)\times(\varphi_1,\varphi_2)$ (not $L_{2,\varphi}$!) formed from the trigonometric function

$$\sin(m\theta)\sin\left[\frac{2\pi s(\varphi-\varphi_1)}{(\varphi_2-\varphi_1)}\right], \qquad \sin(m\theta)\cos\left[\frac{2\pi s(\varphi-\varphi_1)}{\varphi_2-\varphi_1}\right],$$

$$\cos(m\theta)\sin\left[\frac{2\pi s(\varphi-\varphi_1)}{\varphi_2-\varphi_1}\right] \quad \text{and} \quad \cos(m\theta)\cos\left[\frac{2\pi s(\varphi-\varphi_1)}{\varphi_2-\varphi_1}\right],$$

where m and s are nonnegative integers. Let $y(\varphi)\in c^2[\varphi_1,\varphi_2]$ be a function such that

$$y(\varphi) > 0 \quad \text{for } \varphi\in(\varphi_1,\varphi_2),$$

$$y(\varphi_1) = y'(\varphi_1) = y(\varphi_2) = y'(\varphi_2) = 0.$$

Consider functions $\{y(\varphi)\tilde{\phi}_n(\theta,\varphi)\}$. They are linearly independent and satisfy conditions (1.26b). Furthermore, this is a complete set of functions in $L_{2,\varphi}$. Therefore, applying the standard orthogonalization process to this set we obtain the desired orthonormal basis $\{\phi_n\}$.

Choose an orthonormal basis $\{Q_n(r)\}_{n=1}^{\infty}$ in $H_0^2(r_0,R)$ such that all functions Q_n are *piecewise analytic* in (r_0,R) and $Q_n\in C^2[r_0,R]$. Finally, let $\{S_n(t)\}_{n=1}^{\infty}$ be an orthonormal basis in $H^2(0,T_1)$ such that all functions $S_n(t)$ are *analytic* in $(0,T_1)$ and $S_n\in C^2[0,T_1]$. Let Σ be a set of products

$$\Sigma = \{\phi_{n_1}(\theta,\varphi)Q_{n_2}(r)S_{n_3}(t)\}_{n_1,n_2,n_3=1}^{\infty}$$

Then $\Sigma\subset(H_0^2(D_{T_1})\cap\tilde{H}_0^2(D_{T_1}))$ and Σ is an orthonormal basis in $\tilde{H}_0^2(D_{T_1})$. Choose an arbitrary positive integer N and denote

$$m = N^3,$$

$$P_N = \left\{ f\in\tilde{H}_0^2(D_{T_1})\colon f(x,t) = \sum_{n_1,n_2,n_3=1}^{N} c_n\phi_{n_1}(\theta,\varphi)Q_{n_2}(r)S_{n_3}(t), n=(n_1,n_2,n_3)\right\}.$$

We note that $P_N\subset(H_0^2(D_{T_1})\cap\tilde{H}_0^2(D_{T_1}))$. If a function is $f\in P_N$, then $c=c(f)\in\mathbb{R}^m$ will denote the vector of the Fourier coefficients of this function.

In the sequel, we assume that in (1.25) $w\in P_N$.

Consider the function

$$f\in H_0^2(D_{T_1})\cap\tilde{H}_0^2(D_{T_1}).$$

Then

$$\|f\|_{\tilde{H}_0^2(D_{T_1})} \leq \|f\|_{H^2(D_{T_1})}.$$

Furthermore, since all norms are equivalent in finite-dimensional spaces, the following lemma is valid.

LEMMA 1.2. *Norms in $H^2(D_{T_1})$ and $\tilde{H}_0^2(D_{T_1})$ are equivalent for $f \in P_N$.*

Now let us discuss the *noise* issue. If the function ψ in (0.2) is given with a noise, then this noise generates a noise in the functions ψ_3, ψ_4, and, consequently, in the function p because of (1.26a). On the other hand, if the function ψ is given without noise, then generally one should not expect noise in the function p. Because of (1.26a), (1.26b), and Lemma 1.1 we will consider the function p as a *given data* for our inverse problem.

DEFINITION. *We will say that the functions $p \in H^2(D_{T_1})$ is an exact data (generated by the function $a(x)$ in (0.1)) if p satisfies* (1.26a) *and there exists a function $w \in P_N$ such that the function $v = w + p$ satisfies* (1.21) *and* (1.22).

Therefore, in the case of exact data one can find the "target" function $a(x)$ by (1.24) exactly. In the case when p is *not* an exact data we will assume that it can be presented in the form

$$p = p_e + p_\delta \tag{1.27}$$

with $p_e, p_\delta \in H^2(D_{T_1})$. Here function p_e is an exact data, function p_δ corresponds to the noise, and

$$\|p_\delta\|_{H^2(D_{T_1})} \leq \delta, \tag{1.28}$$

where a sufficiently small positive constant δ represents the level of noise. By Tikhonov's principle, in the case of noise data, one should find a sequence $\{w_\delta\} \subset P_N$ such that

$$\lim_{\delta \to 0} \|w_\delta - w_*\|_{H^2(D_{T_1})} = 0, \tag{1.29}$$

where w_* is a solution corresponding to the exact data; see [23] (below we will show that the solution w_* is unique; see Theorem 1.4).

Further, by Tikhonov's principle we can assume that function w_* belongs to an a priori given compact set in $\tilde{H}_0^2(D_{T_1})$. Hence for the sake of definiteness and without loss of generality we will assume that

$$\|w_*\|_{\tilde{H}_0^2(D_{T_1})} < \tfrac{1}{2}. \tag{1.30}$$

Thus

$$c_* \in B_{\frac{1}{2}} \tag{1.31}$$

Denote

$$P_N' = \{w \in P_N : \|w\|_{\tilde{H}_0^2(D_{T_1})} < 1\}.$$

Hence (1.30) implies, in particular, that $w_* \in P_N'$.

Since $T > 3R$ and $T_1 = T - R$, then $R/T_1 < \frac{1}{2}$. Choose a constant α,

$$\alpha \in \left(\frac{R}{T_1}, \frac{1}{2}\right), \tag{1.32}$$

and introduce a function β,

$$\beta(x,t) = \beta(r,t) = -(r + \alpha t). \tag{1.33}$$

Likewise, introduce the domain G,

$$G = \left\{(x,t)\colon \beta(r,t) > -R, r > r_0, \sin\varphi > \frac{r_0}{R}, t > 0\right\}. \tag{1.34}$$

Hence

$$\bar{G} \cap \{r = r_0\} = \left\{(x,t)\colon |x| = r_0, 0 \le t \le \frac{R - r_0}{\alpha}, \sin\varphi > \frac{r_0}{R}\right\}.$$

Consequently (1.32) and (1.34) imply $G \subset D_{T_1}$.

Let λ be a positive parameter. Consider a *cost* functional $J_\lambda(w)$,

$$J_\lambda(w) = \int_G [L(w+p)]^2 e^{2\lambda\beta}\, dx\, dt, \qquad w \in P'_N. \tag{1.35}$$

This functional actually depends on an m-dimensional vector $c = c(w) \in B_1$. Therefore in the cases where we wish to emphasize this dependence we will write $F_\lambda(c)$ instead of $J_\lambda(w)$, keeping in mind, of course, that $F_\lambda(c(w)) = J_\lambda(w)$ and that all the statements about $J_\lambda(w)$ can obviously be reformulated for $F_\lambda(c)$. Now (1.21)–(1.23), (1.30), and (1.31) imply that in the case of *exact* data, our inverse problem can be reformulated as follows.

Minimize functional $J_\lambda(w)$ for an appropriate choice of the parameter λ.

Since $J_\lambda(w) \ge 0$ and $J_\lambda(w_*) = 0$, then w_* is a point of global minimum of $J_\lambda(w)$ (in the case of exact data). Let $J'_\lambda(w)(h), h \in P_N$ be the Fréchet derivative of $J_\lambda(w)$ at the point w. The main result of this paper is as follows.

THEOREM 1.1. *There exists a positive constant $\tilde{\lambda} = \tilde{\lambda}(D_{T_1}, P'_N, \alpha)$ such that for all $\lambda \ge \tilde{\lambda}$ functional $J_\lambda(w)$ is uniformly strictly convex on P'_N. That is,*

$$J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \ge \frac{k}{2}\|h\|^2_{H^2(D_{T_1})}$$

for all $w, h \in P'_N$, where a positive constant $k = k(D_{T_1}, P'_N, \alpha, \lambda)$ does not depend on w, h.

Remark. As mentioned above, the main point of Theorem 1.1 is that beginning with a function, which is a finite-dimensional perturbation of a true solution of the nonlinear Cauchy problem (1.21)–(1.23), one can recover the true solution by a convex minimization procedure. We recall that by Lemma 1.1, Cauchy problem (1.21)–(1.23) is equivalent to the original inverse problem (0.1) and (0.2).

Theorem 1.1 guarantees the global convergence on B_1 of a number of well-known minimization algorithms, provided that the starting point of the iterative process would be the center of B_1. For example, consider the simplest version of the gradient method of the minimization of $F_\lambda(c)$. Fix a number $\lambda \ge \tilde{\lambda}$. Let $H_\lambda(a)$ be the Hessian of $F_\lambda(c)$. Clearly, there exists a positive constant $K = K(\lambda, P'_N)$ such that $y^* H_\lambda(c) y \le K|y|^2$ for all vectors $y \in \mathbb{R}^m$ and for all $a \in B_1$. Theorem 1.1, however, implies that in *addition* to that,

$$k|y|^2 \le y^* H_\lambda(c) y \le K|y|^2, \quad y \in \mathbb{R}^m, \quad a \in P'_N. \tag{1.36}$$

Choose a number ξ such that

$$0 < \xi < \frac{2}{K}. \tag{1.37}$$

Define a sequence $\{a_n\}$ by

$$c_0 = 0,$$

$$c_n = c_{n-1} - \xi \nabla [F_\lambda(c_{n-1})], \qquad n \geq 1. \tag{1.38}$$

Consider a number q:

$$q = \max\{|1 - \xi k|, |1 - \xi K|\}. \tag{1.39}$$

Then

$$0 < q < 1. \tag{1.40}$$

THEOREM 1.2. *Assume that we deal with exact data $p(x,t)$ and fix a number $\lambda \geq \tilde{\lambda}$. Let c_* be the solution of the minimization problem* (1.35) *on B_1 and $c_* \in B_{1/2}$. Then we have $c_n \in B_1$ for all n and*

$$|c_n - c_*| \leq |0 - c_*| q^n < \tfrac{1}{2} q^n.$$

Now consider the case of noisy data.

THEOREM 1.3. *Assume that we deal with noisy data $p(x,t)$ given in the form* (1.27) *and* (1.28) *and fix a number $\lambda \geq \tilde{\lambda}$. Let c_* be the solution of the minimization problem* (1.35) *on B_1 with the exact data and $c_* \in B_{1/2}$. Then there exists a sufficiently small number $\hat{\delta} > 0$ such that if $\delta \leq \hat{\delta}$, then*

$$|c_n - c_*| \leq \frac{1}{2} q^n + C_1 \frac{\xi(1 - q^n)}{1 - q} \delta,$$

where positive constants $C_1, \hat{\delta}$ depend only on D_{T_1}, P'_N, α, and λ.

In particular, for every $\delta \in (0, \hat{\delta})$ choose an integer $n_0 = n_0(\delta)$ such that

$$\frac{1}{2} q^{n_0} < \frac{C_1 \xi}{1 - q} \delta$$

and denote $c^\delta = c_{n_0}$. Then

$$|c^\delta - c_*| < \frac{C_1 \xi}{1 - q} \delta,$$

which corresponds to Tikhonov's principle (1.29).

Thus Theorems 1.2 and 1.3 actually claim the global convergence of the gradient method (1.37) and (1.38) on B_1. Similar convergence results of other versions of the gradient method can be obtained as well. Theorem 1.2, or at least its different versions, is well known, of course, as an almost immediate consequence of (1.36); cf. Hestenes [4]. However, we will give a proof of this theorem since we will need it for the proof of Theorem 1.3.

Note that one can consider the following set $P_N(q)$ instead of P'_N:

$$P_N(q) = \{w \in P_N : \|w\|_{\tilde{H}^2(D_{T_1})} < q\},$$

with an arbitrary positive constant q. Definitely Theorems 1.1–1.3 can be reformulated for this case. Hence the following uniqueness result follows immediately from Theorem 1.2.

THEOREM 1.4. *The Cauchy problem* (1.21)–(1.23) *has at most one solution* $v(x,t)$ *of the form* (1.25), *where function* $w \in P_N$ *and the function* p *satisfies* (1.26a). *In particular, inverse problem* (0.1) *and* (0.2) *has at most one solution such that the function* v *satisfies these conditions.*

Now let us *heuristically* discuss possible implications of these results for *practical* computations. One of the advantages of the functional J_λ is that one should calculate four-dimensional integrals over the domain G only once. In (1.35), a finite number of integrals are calculated, using certain combinations of basic functions, $e^{2\lambda\beta}$ and p. Besides, one should not calculate solutions of the forward problem as soon as data $p(x,t)$ is given, which is convenient since computationally the forward problems are usually time consuming.

We note that the weight function $e^{2\lambda\beta}$ decreases rapidly with respect to both r and t. Generally one should probably not expect that by solving the minimization problem, one could get a reasonable approximation for the solution w_* for *all* $(x,t) \in D_{T_1}$. Instead, one should expect to get a good approximation only in a certain neighborhood of the surface η. And this neighborhood is *not* necessarily small.

Thus by virtue of (1.24) one could get a reasonable approximation for the function $a(x)$ in a certain neighborhood of the surface $\eta \cap \{t = 0\} \subset \mathbb{R}^3$, say in the domain $\{x : r_0 < |x| < r_0 + b, \sin\varphi > (r_0/R)\}$, where b is a positive constant. Then by (1.10)–(1.12) one could solve the following linear time-like Cauchy problem:

$$\Delta\tilde{u} - 2\tilde{u}_{rt} + a(x)\tilde{u} = 0, \tag{1.41}$$

in

$$\left\{(x,t) : r_0 < |x| < r_0 + b, \sin\varphi > \frac{r_0}{R}, t \in (0,T_1)\right\}, \tag{1.42}$$

$$\tilde{u}|_{t=0} = \frac{1}{4\pi|x|}, \tag{1.43}$$

$$\tilde{u}|_\eta = \hat{\psi}_1(x,t+r)|_\eta, \qquad \left.\frac{\partial\tilde{u}}{\partial n}\right|_\eta = \hat{\psi}_2(x,t+r)|_\eta. \tag{1.44}$$

Therefore one can get functions

$$\tilde{u}|_{|x|=r_0+b}, \quad \left.\frac{\partial\tilde{u}}{\partial r}\right|_{|x|=r_0+b}, \quad \text{for } \sin\varphi > \frac{r_0}{R}, \quad t \in (0,T_1),$$

in this way. Then one should take $r_1 = r_0 + b$, etc. This procedure is actually very similar to the layer stripping method; cf. Somersalo, Cheney, Isaakson, and Isaakson [20] and Sylvester [21].

Note that in problem (1.41)–(1.44), Cauchy data are given on the part η of the side of the time cylinder (1.41) rather than on the whole lateral surface, which is

$$\eta \cup \left\{ |x| = r_0 + b, \sin\varphi > \frac{r_0}{R}, t \in (0, T_1) \right\}.$$

This indicates that the Cauchy problem (1.41)–(1.44) is perhaps rather unstable; cf. John [10]. On the other hand, one might expect that the initial data (1.43) might provide "more stability" for this problem. Thus the stability question needs to be investigated further in this case. Also, we note that a numerical method for a similar Cauchy problem with the data on the whole lateral surface has been developed in [13] and [14].

On the other hand, since in an imaging process one is expected to measure scattering data *all around* a bounded medium, then it is reasonable to assume that, in addition, function $a(x)$ is given for $|x| > R$ and function $u(x,t)$ is given at the sphere $\{|x| = R\}$ as well. In this case one should replace in (1.26a) the surface η with the surface

$$\eta_1 = \eta \cup \left\{ |x| = R, \sin\varphi > \frac{r_0}{R}, t \in (0, T_1) \right\}.$$

Thus by (1.25) one would have $w|_{\eta_1} = (\partial w/\partial n)|_{\eta_1} = 0$ and functions $Q_n(r)$ satisfying the following additional conditions $Q_n(R) = Q_n'(R) = 0$. Furthermore, domain G should be replaced with

$$\tilde{G} = \left\{ (x,t) \colon r + \alpha t < R', \, r_0 < r < R, \, \sin\varphi > \frac{r_0}{R}, \, t \in (0, T_1) \right\},$$

where R' is a number such that $R' > R$. Thus $G \subset \tilde{G}$.

An obvious modification of the Theorem 1.1 is valid in this case, of course. Furthermore, we feel that this modified ISP is, by its genuine nature, much more stable than the original one, since we "restrict" function w at $\{|x| = R\}$ as well. Certain indications of this were given in Kazemi and Klibanov [11], Klibanov and Malinsky [13], Komornik and Zuazua [16], and Lasiecka and Triggiani [17], where Lipschitz stability estimates were obtained for some hyperbolic Cauchy problems with the data on the lateral side of the time cylinder (see also references in [11], [13], [16], and [17]). Numerical experiments for these Cauchy problems were performed in Klibanov and Rakesh [14], and indeed they demonstrated high stability of these problems. Thus in this case one might expect to have a good approximation for $w_*(x,t)$ for all $(x,t) \in D_{T_1}$ by minimization of J_λ. This line of development will be discussed elsewhere.

2. Auxiliary lemma. The main result of this section is Lemma 2.3, which is based on the special properties of the finite-dimensional space P_N. Since $Q_n(r)$ are piecewise analytic functions then the following lemma is true.

LEMMA 2.1. *There exists a number $R_1 = R_1(N) \in (0, R)$ such that if*

$$\sum_{n=1}^{N} d_n Q_n(r) = 0 \quad \text{for } r \in (0, R_1),$$

with some constants d_n, then

$$\sum_{n=1}^{N} d_n Q_n(r) \equiv 0 \text{ in } (0, R),$$

where functions Q_n were introduced in §2.

LEMMA 2.2. *Let $R_1 = R_1(N)$ be the number defined in the Lemma* 2.1 *and $R_2 \in [R_1, R)$. Let E be the time cylinder such that*

$$E = \left\{(x,t)\colon r_0 < |x| < R_2, \sin\varphi > \frac{r_0}{R}, t \in (t_1,t_2), \text{ where } 0 \le t_1 < t_2 \le T_1\right\}.$$

Then there exists a positive constant $\gamma_1 = \gamma_1(E, D_{T_1}, P_N)$ such that

$$\|f\|_{H^1(E)} \ge \gamma_1 \|f\|_{\tilde{H}_0^2(D_{T_1})}, \qquad f \in P_N.$$

Proof. Since function $f \in P_N$ are analytic in respect to $t \in (0,T)$ then by Lemma 2.1 the following implication is valid:

$$\{f = 0 \text{ in } E\} \to \{f = 0 \text{ in } D_{T_1}\}.$$

Therefore for $f \in P_N$ norms of $H^1(E)$ and $\tilde{H}_0^2(D_{T_1})$ are equivalent since P_N is a finite-dimensional space. □

Finally, Lemmas 1.2 and 2.2 imply that the following lemma is true.

LEMMA 2.3. *Let the number R_2 and the cylinder E be the same as in Lemma* 2.3. *Then there exists a positive constant $\gamma = \gamma(E, D_{T_1}, P_N)$ such that*

$$\|f\|_{H^1(E)} \ge \gamma \|f\|_{H^2(D_{T_1})}, \qquad f \in P_N.$$

3. Special form of a Carleman estimate for P_N. Let us remember by (1.23) and (1.32)–(1.33) we have the following relations:

$$L_0 u = \Delta u - 2u_{rt}, \tag{3.1}$$

$$\beta(r,t) = -(r + \alpha t), \qquad \alpha = \text{ constant } \in \left(\frac{R}{T_1}, \frac{1}{2}\right), \tag{3.2}$$

$$G = \left\{(x,t)\colon \beta(r,t) > -R, r > r_0, \sin\varphi > \frac{r_0}{R}, t > 0\right\}. \tag{3.3}$$

Hence the boundary ∂G of the domain G consists from exactly four pieces, $\omega_1, \omega_2, \omega_3$, and ω_4:

$$\omega_1 = \left\{r = r_0, \sin\varphi > \frac{r_0}{R}, 0 < t < \frac{R - r_0}{\alpha} < T_1\right\}, \tag{3.4}$$

$$\omega_2 = \left\{t = 0, r_0 < r < R, \sin\varphi > \frac{r_0}{R}\right\}, \tag{3.5}$$

$$\omega_3 = \left\{r + \alpha t = R, r > r_0, \sin\varphi > \frac{r_0}{R}, t > 0\right\}, \tag{3.6}$$

$$\omega_4 = \left\{\sin\varphi = \frac{r_0}{R}, \beta(r,t) > -R, r > r_0, t > 0\right\}, \tag{3.7}$$

$$\partial G = \omega_1 \cup \omega_2 \cup \omega_3 \cup \omega_4.$$

In the sequel, C will denote different positive constants dependent only on G, P_N, and $\alpha, C = C(G, P_N, \alpha) > 1$. However, in Lemmas 3.1–3.6, C does not depend on P_N. The main result of this section is the following theorem.

THEOREM 3.1 (special form of a Carleman estimate for P_N). *There exists a sufficiently large constant $\tilde{\lambda} = \tilde{\lambda}(G, P_N, \alpha)$ such that for all $\lambda \geq \tilde{\lambda}$ and for all functions $u \in P_N$ the following Carleman estimate holds:*

$$\int_G (L_0 u)^2 e^{2\lambda\beta}\, dx\, dt \geq C\lambda \int_G (|\nabla u|^2 + u_t^2) e^{2\lambda\beta}\, dx\, dt - Ce^{-2\lambda R} \int_{\omega_3} (|\nabla u|^2 + u_t^2)\, d\sigma.$$

Proof of this theorem consists of proofs of several lemmas (see Lemmas 3.1–3.7 below). Proofs of these lemmas are rather simple, but they require a relatively large number of rather routine calculations, which is "commonplace" in the theory of Carleman estimates (cf. Lavrent'ev et al. [18], Chapter 4). A more elegant proof would probably be the symbol method, as it was done, for instance, by Hörmander [8] and Isàkov [9] for "standard" Carleman estimates, that is, for estimates which are valid for the *whole* Sobolev spaces. But we do not know how to use this method in our *particular* case, because our estimate is valid only for *finite-dimensional* space P_N, and because of a major difficulty in our particular case described below.

Now let us explain why we cannot use the standard Carleman estimate here. First of all, let us remember briefly the Carleman estimate for the operator $Pu = u_{tt} - \Delta u$; cf. Isàkov [9] and Lavrent'ev et al. [18]. Consider domain

$$\Omega = \{|x| < 1, t \in (-b, b)\},$$

where $b =$ constant > 0. Consider a weight function $\beta_1(x,t) = |x|^2 - \sigma t^2$, where $\sigma =$ constant $\in (0,1)$. Choose a constant c such that

$$c \in (0,1) \quad \text{and} \quad \sqrt{\frac{1-c}{\sigma}} < b.$$

Consider domain

$$\Omega_c = \{(x,t)\colon |x| < 1, \beta_1(x,t) > c\}.$$

Then $\Omega_c \subset \Omega$ and the following Carleman estimate is valid:

$$\begin{aligned} \int_{\Omega_c} (Pu)^2 e^{2\lambda\beta_1}\, dx\, dt \geq & D\lambda \int_{\Omega_c} (|_{\Omega_c}(|\nabla u|^2 + u_t^2) e^{2\lambda\beta_1}\, dx\, dt \\ & + D\lambda^3 \int_{\Omega_c} u^2 e^{2\lambda\beta_1}\, dx\, dt, \end{aligned} \tag{3.8}$$

for all $u \in \overset{\circ}{H}{}^2(\Omega_c)$ and for all $\lambda \geq \lambda_0$, where positive constants D, λ_0 depend only on Ω, Ω_c, and σ.

In particular, the last estimate implies, by $b > 1$, uniqueness of a time-like Cauchy problem

$$|u_{tt} - \Delta u| \leq A(|\nabla u| + |u_t| + |u|) \quad \text{in } \Omega,$$

$$u|_{|x|=1} = \left.\frac{\partial u}{\partial r}\right|_{|x|=1} = 0, \tag{3.9}$$

where $A =$ constant > 0; see Hörmander [8], Isàkov [9], and Lavrent'ev et al. [18]. Likewise, using Carleman estimate (3.8) and prescribing nonzero Cauchy data at the surface $\{|x| = 1\}$ one can get Hölder and even Lipshitz stability estimates for this Cauchy problem; cf. Kazemi and Klibanov [11], and Klibanov and Malinsky [13].

Now note that the boundary $\partial\Omega_c$ of the domain of integration in (3.8) consists of *exactly* two parts $\partial_1\Omega_c, \partial_2\Omega_c$, where

$$\partial_1\Omega_c = \{\beta_1(x,t) = c\}, \qquad \partial_2\Omega_c = \{|x| = 1\} \cap \bar{\Omega}, \tag{3.10}$$
$$\partial\Omega = \partial_1\Omega_c \cup \partial_2\Omega_c.$$

Thus $\partial_1\Omega_c$ is the *level surface* of the weight function and $\partial_2\Omega_c$ is the surface, where *Cauchy data* are given.

Property (3.10) *is a genuine feature of the Carleman estimates theory.*

Remark. Sometimes domains of integrations with different properties can be considered. However, as soon as one needs to get a uniqueness or a stability result (on the basis of Carleman estimates) one must inevitably employ domains with the property (3.10).

In the case of Theorem 3.1, however, surface ω_2 is a part of the boundary ∂G; see (3.4)–(3.7). But definitely ω_2 is neither a level surface of the weight function nor a surface where Cauchy data are given (they are given on ω_1 and ω_4, since $u \in P_N \subset H_0^2(D_{T_1})$). All the other surfaces in (3.4)–(3.7) are either level surfaces of the weight function or surfaces, where Cauchy data are given. Nevertheless the surface ω_2 *has* to be a part of the boundary of the domain of integration, because of our need to estimate the integral

$$\int_G e^{2\lambda\beta} \left(\int_0^t u(x,\tau)\, d\tau \right)^2 dx\, dt, \qquad u \in L_2(G)$$

through the integral

$$\int_G u^2 e^{2\lambda\beta}\, dx\, dt,$$

whichever domain G would be (see Lemma 3.8 and (1.22)). Therefore the presence of the surface ω_2 is the major difficulty of our particular case of the Carleman estimate.

Remark. In the sequel, we will always use the fact that in the domain $\{r > r_0, \sin\varphi > r_0/R\}$

$$A_2 \sum_{i=1}^{3} u_{x_i}^2 \le (u_r^2 + u_\theta^2 + u_\varphi^2) \le A_2 \sum_{i=1}^{3} u_{x_i}^2,$$

for all smooth functions u, where positive constants A_1, A_2 depend only on r_0 and r_0/R.

Below we prove Theorem 3.1 by proving Lemmas 3.1–3.7. In these lemmas we first obtain pointwise differential identities using the rule of differentiation of products. Then by integrating these identities over the domain G and employing Gauss' formula, we will get desired L_2 estimates.

Thus in Lemmas 3.1 and 3.2 we estimate from below the integral

$$\int_G 2u_r L_0 u e^{2\lambda\beta}\, dx\, dt.$$

Then in Lemmas 3.3 and 3.4 we estimate from below the integral

$$\int_G -2u_t L_0 u e^{2\lambda\beta}\,dx\,dt.$$

Finally, a combination of two latter estimates and properties of the set P_N will provide the desired Carleman estimate, that is, an estimate from below the integral

$$\int_G (L_0 u)^2 e^{2\lambda\beta}\,dx\,dt, \qquad u \in P_N;$$

see Lemmas 3.5–3.7.

LEMMA 3.1. *For all functions $u \in C^2(\bar{G})$ the following identity is valid:*

(3.11)
$$(2u_r L_0 u) r^2 \sin\varphi e^{2\lambda\beta} = 2\lambda(1-2\alpha)u_r^2 r^2 \sin\varphi e^{2\lambda\beta} - 2\lambda\left(\frac{u_\theta^2}{\sin^2\varphi} + u_\varphi^2\right)\sin\varphi e^{2\lambda\beta} + V_t + \frac{\partial U_1}{\partial r} + \frac{\partial U_2}{\partial\theta} + \frac{\partial U_3}{\partial\varphi},$$

where

(3.12)
$$V = -2u_r^2 r^2 \sin\varphi e^{2\lambda\beta}.$$

The vector functions $U = (U_1, U_2, U_3)$ belongs to C^1 and satisfies the estimate

(3.13)
$$|U| \le C|\nabla u|^2 e^{2\lambda\beta}$$

and

(3.14)
$$U_2(r, 0, \varphi, t) = U_2(r, 2\pi, \varphi, t).$$

Proof. The left-hand side of (3.11) can be rewritten as

$$2u_r\left[u_{rr} + \frac{2}{r}u_r + \frac{1}{r^2\sin^2\varphi}u_{\theta\theta} + \frac{1}{r^2\sin\varphi}\frac{\partial}{\partial\varphi}(\sin\varphi u_\varphi) - 2u_{rt}\right] \times r^2\sin\varphi e^{-2\lambda(r+\alpha t)}.$$

Hence by (3.2)

$$\begin{aligned}
2u_r(L_0u)r^2\sin\varphi e^{2\lambda\beta} &= \frac{\partial}{\partial r}(u_r^2 r^2\sin\varphi e^{2\lambda\beta}) + 2\lambda r^2\sin\varphi u_r^2 e^{2\lambda\beta} \\
&\quad + \frac{\partial}{\partial\theta}\left(2u_r u_\theta\frac{\sin\varphi}{\sin^2\varphi}e^{2\lambda\beta}\right) - 2u_{r\theta}u_\theta\frac{\sin\varphi}{\sin^2\varphi}e^{2\lambda\beta} \\
&\quad + \frac{\partial}{\partial\varphi}(2u_r u_\varphi\sin\varphi e^{2\lambda\beta}) - 2u_{r\varphi}u_\varphi\sin\varphi e^{2\lambda\beta} \\
&\quad + \frac{\partial}{\partial t}(-2u_r^2 r^2\sin\varphi e^{2\lambda\beta}) - 4\lambda\alpha u_r^2 r^2\sin\varphi e^{2\lambda\beta} \\
&= 2\lambda(1-2\alpha)u_r^2 r^2\sin\varphi e^{2\lambda\beta} - 2\lambda\left(\frac{u_\theta^2}{\sin_\varphi^2} + u_\varphi^2\right)\sin\varphi e^{2\lambda\beta} \\
&\quad + V_t + \frac{\partial U_1}{\partial r} + \frac{\partial U_2}{\partial\theta} + \frac{\partial U_3}{\partial\varphi},
\end{aligned}$$

where the function V is as in (3.12), and functions U_1, U_2, U_3 have the forms

$$U_1 = \left(u_r^2 r^2 - \frac{u_\theta^2}{\sin_\varphi^2} - u_\varphi^2\right) \sin\varphi e^{2\lambda\beta}$$

$$U_2 = 2u_r u_\theta \frac{\sin\varphi}{\sin^2\varphi} e^{2\lambda\beta}, \qquad U_3 = 2u_r u_\varphi \sin\varphi e^{2\lambda\beta}. \qquad \square$$

LEMMA 3.2. *For all positive λ and for all functions $u \in H_0^2(G)$ the following estimate is valid*:

$$\begin{aligned} \int_G 2u_r L_0 u e^{2\lambda\beta}\, dx\, dt &\geq 2\lambda \int_G [(1-2\alpha)u_r^2 - Cu_\theta^2 - u_\varphi^2] e^{2\lambda\beta}\, dx\, dt \\ &\quad + \int_{\omega_2} 2u_r^2 e^{2\lambda\beta}\, dx - Ce^{-2\lambda R} \int_{\omega_3} |\nabla u|^2\, d\sigma. \end{aligned} \tag{3.15}$$

Proof. Since $\alpha \in (0, \frac{1}{2})$ then $1 - 2\alpha > 0$. Let $u \in C^2(\bar{G}) \cap H_0^2(G)$ in (3.11). Integrate (3.11) over G using Gauss' formula and (3.12)–(3.14). Noting that $\beta|_{\omega_3} = -R$ and $-(1/\sin^2\varphi) > -(R/r_0)^2$ in G we obtain (3.15) for all functions $u \in C^2(\bar{G}) \cap H_0^2(G)$. But since this set of functions is dense in $H_0^2(G)$, then (3.15) also holds for all $u \in H_0^2(G)$. $\square$

LEMMA 3.3. *For all functions $u \in C^2(\bar{G})$ the following identity is valid*:

$$\begin{aligned} &[-2u_t(L_0 u) r^2 \sin\varphi] e^{2\lambda\beta} \\ &\quad = 4\lambda[u_t^2 + \frac{\alpha}{2} u_r^2 - u_t u_r] r^2 \sin\varphi e^{2\lambda\beta} \\ &\qquad + 2\lambda\alpha \left(\frac{u_\theta^2}{\sin^2\varphi} + u_\varphi^2\right) \sin\varphi e^{2\lambda\beta} - 4u_t^2 r \sin\varphi e^{2\lambda\beta} \\ &\qquad + W_t + \frac{\partial Z_1}{\partial r} + \frac{\partial Z_2}{\partial\theta} + \frac{\partial Z_3}{\partial\varphi}, \end{aligned} \tag{3.16}$$

where

$$W = \left(u_r^2 r^2 + \frac{u_\theta^2}{\sin^2\varphi} + u_\varphi^2\right) \sin\varphi e^{2\lambda\beta}. \tag{3.17}$$

the vector function $Z = (Z_1, Z_2, Z_3)$ belongs to C^1, satisfies the estimate

$$|Z| \leq C(|\nabla u|^2 + u_t^2) e^{2\lambda\beta}, \tag{3.18}$$

and

$$Z_2(r, 0, \varphi, t) = Z_2(r, 2\pi, \varphi, t). \tag{3.19}$$

Proof. The left-hand side of (3.16) can be rewritten as

$$\begin{aligned} &-2u_t(L_0 u) r^2 \sin\varphi e^{2\lambda\beta} \\ &\quad = -2u_t \left[u_{rr} + \frac{2}{r} u_r + \frac{1}{r^2 \sin^2\varphi} u_{\theta\theta} + \frac{1}{r^2 \sin\varphi} \frac{\partial}{\partial\varphi} (\sin\varphi u_\varphi) - 2u_{rt}\right] r^2 \sin\varphi e^{2\lambda\beta}. \end{aligned}$$

Hence

$$
\begin{aligned}
-2u_t(L_0u)r^2\sin\varphi e^{2\lambda\beta} &= \frac{\partial}{\partial r}\left(-2u_tu_rr^2\sin\varphi e^{2\lambda\beta}\right)\\
&+ 2u_{rt}u_rr^2\sin\varphi e^{2\lambda\beta} + 4u_tu_rr\sin\varphi e^{2\lambda\beta}\\
&- 4\lambda u_tu_rr^2\sin\varphi e^{2\lambda\beta} - 4u_tu_rr\sin\varphi e^{2\lambda\beta}\\
&+ \frac{\partial}{\partial\theta}\left(-2u_tu_\theta\frac{\sin\varphi}{\sin^2\varphi}e^{2\lambda\beta}\right) + 2u_{t\theta}u_\theta\frac{\sin\varphi}{\sin^2\varphi}e^{2\lambda\beta}\\
&+ \frac{\partial}{\partial\varphi}\left(-2u_tu_\varphi\sin\varphi e^{2\lambda\beta}\right) + 2u_{t\varphi}u_\varphi\sin\varphi e^{2\lambda\beta}\\
&+ \frac{\partial}{\partial r}\left(2u_t^2r^2\sin\varphi e^{2\lambda\beta}\right) + 4\lambda u_t^2r^2\sin\varphi e^{2\lambda\beta} - 4u_t^2r\sin\varphi e^{2\lambda\beta}\\
&= 4\lambda\left[u_t^2 + \frac{\alpha}{2}u_r^2 - u_tu_r\right]r^2\sin\varphi e^{2\lambda\beta}\\
&+ 2\lambda\alpha\left(\frac{u_\theta^2}{\sin^2\varphi} + u_\varphi^2\right)\sin\varphi e^{2\lambda\beta} - 4u_t^2r\sin\varphi e^{2\lambda\beta}\\
&+ W_t + \frac{\partial Z_1}{\partial r} + \frac{\partial Z_2}{\partial\theta} + \frac{\partial Z_3}{\partial\varphi},
\end{aligned}
$$

where

$$
W = \left(u_r^2r^2 + \frac{u_\theta^2}{\sin^2\varphi} + u_\varphi^2\right)\sin\varphi e^{2\lambda\beta},
$$

and components of the vector function Z have the forms

$$
Z_1 = 2(u_t^2 - u_tu_r)r^2\sin\varphi e^{2\lambda\beta}
$$

$$
Z_2 = -2u_tu_\theta\frac{\sin\varphi}{\sin^2\varphi}e^{2\lambda\beta}, \qquad Z_3 = -2u_tu_\varphi\sin\varphi e^{2\lambda\beta}. \qquad \square
$$

LEMMA 3.4. *For all positive λ and for all functions $u \in H_0^2(G)$, the following inequality holds*:

$$
\begin{aligned}
-\int_G 2u_tL_0ue^{2\lambda\beta}\,dx\,dt &\geq 4\lambda\int_G\left(u_t^2 + \frac{\alpha}{2}u_r^2 - |u_tu_r|\right)e^{2\lambda\beta}\,dx\,dt\\
&+ 2\lambda\alpha\int_G(u_\theta^2 + u_\varphi^2)e^{2\lambda\beta}\,dx\,dt - \frac{4}{r_0}\int_G u_t^2e^{2\lambda\beta}\,dx\,dt\\
&- \int_{\omega_2}\left[u_r^2 + \frac{1}{r^2}\left(\frac{u_\theta^2}{\sin^2\varphi} + u_\varphi^2\right)\right]e^{2\lambda\beta}\,dx\\
&- Ce^{-2\lambda R}\int_{\omega_3}(|\nabla u|^2 + u_t^2)\,d\sigma.
\end{aligned}
\tag{3.20}
$$

Proof. Integrate (3.16) over G, using Gauss' formula and (3.17)–(3.19). Noting that $\beta|_{\omega_3} = -R$ and $(1/\sin^2\varphi) \geq 1$, we obtain (3.20) for all functions $u \in C^2(\bar{G}) \cap H_0^2(G)$. Since the last set of functions is dense in $H_0^2(G)$, then (3.20) is valid for $u \in H_0^2(G)$ as well. $\square$

LEMMA 3.5. *Choose a constant s such that*

$$
0 < s < \frac{1-2\alpha}{1-\alpha}. \tag{3.21}
$$

Then there exists a constant $\lambda_1 = \lambda_1(r_0, \alpha, R) = \lambda_1(G)$ *such that for all* $\lambda > \lambda_1$ *and for all functions* $u \in H_0^2(G)$ *the following inequality is valid*:

$$\begin{aligned}
&\int_G (2u_r - 2su_t) L_0 u e^{2\lambda\beta}\, dx\, dt \\
&\quad \geq C\lambda \int_G (u_r^2 + u_t^2) e^{2\lambda\beta}\, dx\, dt - C\lambda \int_G (u_\theta^2 + u_\varphi^2) e^{2\lambda\beta}\, dx\, dt \\
&\qquad + \int_{\omega_2} [(2-s)u_r^2 - C(u_\theta^2 + u_\varphi^2)] e^{2\lambda\beta} - Ce^{-2\lambda R} \int_{\omega_3} (|\nabla u|^2 + u_t^2)\, d\sigma.
\end{aligned} \tag{3.22}$$

Proof. Multiply inequality (3.20) by s and add to (3.15). Noting that $s < 1$ and $-(1/r^2 \sin^2 \varphi) \geq -C$ in G, we obtain

$$\begin{aligned}
&\int_G (2u_r - 2su_t) L_0 u e^{2\lambda\beta}\, dx\, dt \\
&\quad \geq 2\lambda \int_G \{(1 - 2\alpha + \alpha s)u_r^2 - 2s|u_t||u_r| + 2su_t^2\} e^{2\lambda\beta}\, dx\, dt \\
&\qquad - C\lambda \int_G (u_\theta^2 + u_\varphi^2) e^{2\lambda\beta}\, dx\, dt - \frac{4s}{r_0} \int_G u_t^2 e^{2\lambda\beta}\, dx\, dt \\
&\qquad + \int_{\omega_2} [(2-s)u_r^2 - C(u_\theta^2 + u_\varphi^2)] e^{2\lambda\beta}\, dx - Ce^{-2\lambda R} \int_{\omega_3} (|\nabla u|^2 + u_t^2)\, d\sigma.
\end{aligned} \tag{3.23}$$

Consider the quadratic form

$$f(x_1, x_2) = (1 - 2\alpha + \alpha s)x_1^2 - 2sx_1x_2 + 2sx_2^2.$$

Note that $1 - 2\alpha + 2s > 0$, since $\alpha \in (0, \frac{1}{2})$. Therefore this quadratic form is positive definite for s satisfying (3.21). Hence there exists a positive constant $b = b(\alpha)$ such that

$$f(x_1, x_2) \geq b(x_1^2 + x_2^2),$$

for all x_1, x_2. Let $\lambda_1 = 4s/br_0$. Hence, for $\lambda > \lambda_1$, (3.23) implies

$$\begin{aligned}
&\int_G (2u_r - 2su_t) L_0 u e^{2\lambda\beta}\, dx\, dt \\
&\quad \geq C\lambda \int_G (u_r^2 + u_t^2) e^{2\lambda\beta}\, dx\, dt - C\lambda \int_G (u_\theta^2 + u_\varphi^2) e^{2\lambda\beta}\, dx\, dt \\
&\qquad + \int_{\omega_2} [(2-s)u_r^2 - C(u_\theta^2 + u_\varphi^2)] e^{2\lambda\beta}\, dx - Ce^{-2\lambda R} \int_{\omega_3} (|\nabla u|^2 + u_t^2)\, d\sigma,
\end{aligned}$$

which is exactly (3.22). □

LEMMA 3.6. *There exists a positive constant* $\lambda_2 = \lambda_2(G)$ *such that for all* $\lambda \geq \lambda_2$ *and for all functions* $u \in H_0^2(G)$ *the following inequality is valid*:

$$\begin{aligned}
\int_G (L_0 u)^2 e^{2\lambda\beta}\, dx\, dt \geq\ & C\lambda \int_G (u_r^2 + u_t^2) e^{2\lambda\beta}\, dx\, dt \\
& - C\lambda \int_G (u_\theta^2 + u_\varphi^2) e^{2\lambda\beta}\, dx\, dt \\
& + \frac{1}{2} \int_{\omega_2} [(2-s)u_r^2 - C(u_\theta^2 + u_\varphi^2)] e^{2\lambda\beta}\, dx \\
& - Ce^{-2\lambda R} \int_{\omega_3} (|\nabla u|^2 + u_t^2)\, d\sigma.
\end{aligned} \tag{3.24}$$

Proof. The Cauchy inequality leads to

$$(2u_r - 2su_t)L_0u \le (L_0u)^2 + u_r^2 + (L_0u)^2 + s^2u_t^2$$
$$\le 2(L_0u)^2 + u_r^2 + u_t^2.$$

Hence (3.24) follows immediately from the last inequality and (3.22). □

Until now we have *not* used the assumption $u \in P_N$. As a price for that we have negative signs at u_θ^2, u_φ^2 in (3.24). In order to get rid of this defect we should "supress negative integrals"

$$-C\lambda \int_G (u_\theta^2 + u_\varphi^2)e^{2\lambda\beta}\,dx\,dt, \qquad -\int_{\omega_2} C(u_\theta^2 + u_\varphi^2)e^{2\lambda\varphi}\,dx$$

by "positive integrals"

$$C\lambda \int_G (u_r^2 + u_t^2)e^{2\lambda\beta}\,dx\,dt, \qquad \int_{\omega_2} u_r^2 e^{2\lambda\beta}\,dx.$$

Thus now is the time to use properties of P_N.

LEMMA 3.7. *For all functions $u \in P_N$ the following estimates are valid*:

$$\int_G (u_\theta^2 + u_\varphi^2)e^{2\lambda\beta}\,dx\,dt \le \frac{C}{\lambda}\int_G u_r^2 e^{2\lambda\beta}\,dx\,dt, \tag{3.25}$$

$$\int_{\omega_2} (u_\theta^2 + u_\varphi^2)e^{2\lambda\beta}\,dx \le \frac{C}{\lambda}\int_{\omega_2} u_r^2 e^{2\lambda\beta}. \tag{3.26}$$

Proof. We will only prove the estimate

$$\int_G u_\theta^2 e^{2\lambda\beta}\,dx\,dt \le \frac{C}{\lambda}\int_G u_r^2 e^{2\lambda\beta}\,dx\,dt \tag{3.27}$$

since proofs of (3.25) and (3.26) are completely parallel.

Consider the function $u \in P_N$. Since functions $\{\phi_{n_1}\}$ form an orthonormal basis in L_2, φ, then the function u can be represented in the form

$$u = \sum_{n_1=1}^{N} u_{n_1}(r,t)\phi_{n_1}(\theta,\varphi), \tag{3.28}$$

where

$$u_{n_1}(r,t) = \int_0^{2\pi} d\theta \int_{\varphi_1}^{\varphi_2} u(r,\theta,\varphi,t)\phi_{n_1}(\theta,\varphi)\sin\varphi\,d\varphi.$$

Hence

$$u_\theta = \sum_{n_1=1}^{N} u_{n_1}(r,t)\,\frac{\partial\phi_{n_1}}{\partial\theta}\,(\theta,\varphi).$$

Hence by the Cauchy inequality,

$$|u_\theta| \le \sqrt{C}\left[\sum_{n_1=1}^{N} u_{n_1}^2(r,t)\right]^{\frac{1}{2}}. \tag{3.29}$$

Denote $q = (R - r_0/\alpha)$. Since the function $\beta = \beta(r,t)$ does not depend on θ, φ then (3.28) and Parceval theorem imply

$$\int_G u^2 e^{2\lambda\beta}\,dx\,dt = \sum_{n_1=1}^{N} \int_0^q dt \int_{r_0}^{R-\alpha t} u_{n_1}^2(r,t) e^{2\lambda\beta} r^2\,dr. \tag{3.30}$$

Similarly (3.29) leads to

$$\int_G u_\theta^2 e^{2\lambda\beta}\,dx\,dt \le C \sum_{n_1=1}^{N} \int_0^q dt \int_{r_0}^{R-\alpha t} u_{n_1}^2(r,t) e^{2\lambda\beta} r^2\,dr.$$

Hence (3.30) and the last estimate imply

$$\int_G u_\theta^2 e^{2\lambda\beta}\,dx\,dt \le C \int_G u^2 e^{2\lambda\beta}\,dx\,dt. \tag{3.31}$$

Now recall that $P_N \subset H_0^2(D_{T_1})$, which means, in particular, that $u|_{r=r_0} = u_r|_{r=r_0} = 0$ for all $u \in P_N$. Hence,

$$u = \int_{r_0}^{r} u_r(y,\theta,\varphi,t)\,dy, \qquad u \in P_N.$$

Hence

$$\begin{aligned}
\int_G u^2 e^{2\lambda\beta}\,dx\,dt &= \int_0^{2\pi} d\theta \int_{\varphi_1}^{\varphi_2} \sin\varphi\,d\varphi \int_0^q dt \int_{r_0}^{R-\alpha t} r^2 e^{2\lambda\beta} \left(\int_{r_0}^{r} u_r\,dy\right)^2 dr \\
&\le R^3 \int_0^{2\pi} d\theta \int_{\varphi_1}^{\varphi_2} \sin\varphi\,d\varphi \int_0^q dt \int_{r_0}^{R-\alpha t} e^{2\lambda\beta} \int_{r_0}^{r} u_r^2\,dy \\
&= R^3 \int_0^{2\pi} d\theta \int_{\varphi_1}^{\varphi_2} \sin\varphi\,d\varphi \int_0^q e^{-2\lambda\alpha t}\,dt \int_{r_0}^{R-\alpha t} u_r^2\,dy \int_y^{R-\alpha t} e^{-2\lambda r}\,dr \\
&= R^3 \int_0^{2\pi} d\theta \int_{\varphi_1}^{\varphi_2} \sin\varphi\,d\varphi \int_0^q e^{-2\lambda\alpha t}\,dt \\
&\quad \times \int_{r_0}^{R-\alpha t} u_r^2\,dy \left[\frac{e^{-2\lambda y} - e^{-2\lambda(R-\alpha t)}}{2\lambda}\right] \\
&\le \frac{R^3}{2\lambda r_0^2} \int_0^{2\pi} d\theta \int_{\varphi_1}^{\varphi_2} \sin\varphi\,d\varphi \int_0^q dt \int_{r_0}^{R-\alpha t} u_r^2 e^{2\lambda\beta(y,t)} y^2\,dy \\
&\le \frac{C}{\lambda} \int_G u_r^2 e^{2\lambda\beta}\,dx\,dt.
\end{aligned}$$

Hence we have actually proven that

$$\int_G u^2 e^{2\lambda\beta}\,dx\,dt \le \frac{C}{\lambda} \int_G u_r^2 e^{2\lambda\beta}\,dx\,dt, \qquad u \in H_0^2(G). \tag{3.32}$$

Finally, (3.31) and (3.32) lead to

$$\int_G u_\theta^2 e^{2\lambda\beta}\,dx\,dt \le \frac{C}{\lambda} \int_G u_r^2 e^{2\lambda\beta}\,dx\,dt, \qquad u \in P_N. \qquad \square$$

Proof of Theorem 3.1. This proof follows immediately from Lemmas 3.6 and 3.7. We close this section with one auxiliary result which is quite similar to Lemma 3.7. But the difference is that this result is valid for *all* $u \in L_2(G)$. We omit the proof since a quite similar result is known; see [12].

LEMMA 3.8. *For all functions* $u \in L_2(G)$ *the following inequality is valid*

$$\int_G \left[\int_0^t u(x,t)\,dt\right]^2 e^{2\lambda\beta}\,dx\,dt \le \frac{D}{\lambda}\int_G u^2 e^{2\lambda\beta}\,dx\,dt,$$

where the positive constant D *depends only on the domain* G *and on the constant* α.

4. Proof of Theorem 1.1. Let us remember that the functional $J_\lambda(w)$ has the form

$$J_\lambda(w) = \int_G [L(p+w)]^2 e^{2\lambda\beta}\,dx\,dt, \qquad w \in P_N'. \tag{4.1}$$

Here function $p \in H^2(D_{T_1})$,

$$P_N' = \{w \in P_N : \|w\|_{\tilde{H}_0^2(D_{T_1})} < 1\},$$

and the nonlinear operator L has the form (see (1.22))

$$\begin{aligned} Lv = {} & L_0 v + 2\nabla v \int_0^t \nabla v(x,\tau)\,d\tau - 2v_t \int_0^t v_r(x,\tau)\,d\tau \\ & - 2v_r v + \frac{2}{r}\,v_t + 2\nabla v \nabla g - 2v_t g'(r), \end{aligned} \tag{4.2}$$

with

$$L_0 v = \Delta v - 2v_{rt}, \qquad g(r) = \ln\frac{1}{4\pi r}.$$

The main idea of the proof is that first we prove, by rather simple estimates, that

$$\begin{aligned} & J_\lambda(w+h) - J_\lambda(w) - J_\lambda'(w)(h) \\ & \ge \frac{1}{4}\int_G (L_0 h)^2 e^{2\lambda\beta}\,dx\,dt \\ & \quad - CM^2 \int_G \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left(\int_0^t |\nabla h(x,\tau)|\,d\tau\right)^2 \right\} e^{2\lambda\beta}\,dx\,dt, \end{aligned}$$

where the positive constant M does not depend on λ, w, and h. Then we note that by Theorem 3.1, (3.32), and Lemma 3.8 the first integral on the right-hand side of the latter estimate, roughly speaking, dominates the second integral. Finally, Lemma 2.4 will provide the desired result.

Proof. First of all, we should calculate

$$J_\lambda(w+h) - J_\lambda(w) \text{ and } J_\lambda'(w)(h)$$

for $w, h \in P_N'$. By (4.1)

$$\begin{aligned} & J_\lambda(w+h) - J_\lambda(w) \\ & = \int_G [L(p+w+h) - L(p+w)][L(p+w+h) + L(p+w)] e^{-2\lambda\beta}\,dx\,dt. \end{aligned} \tag{4.3}$$

By virtue of (4.2),

$$\begin{aligned} L(p+w+h) &= L(p+w) + L_0 h + 2\nabla(p+w)\int_0^t \nabla h(x,\tau)\,d\tau \\ &\quad + 2\nabla h \int_0^t \nabla(p+w)(x,\tau)\,d\tau - 2(p+w)_t \int_0^t h_r(x,\tau)\,d\tau \\ &\quad - 2h_t \int_0^t (p+w)_r(x,\tau)\,d\tau - 2(p+w)h_r \\ &\quad + \frac{2}{r}\,h_t + 2\nabla h\nabla g - 2(p+w)_r h - 2h_t g'(r) \\ &\quad - 2h_r h + 2\nabla h \int_0^t \nabla h(x,\tau)\,d\tau - 2h_t \int_0^t h_r(x,\tau)\,d\tau. \end{aligned}$$

Hence

$$L(p+w+h) - L(p+w) = L_0(h) + L_1(p+w,h) + L_2(h), \tag{4.4}$$

where L_1 is a linear operator with respect to h, and L_2 is a nonlinear operator. Namely,

$$\begin{aligned} L_1(p+w,h) &= 2\nabla(p+w)\int_0^t \nabla h(x,\tau)\,d\tau + 2\nabla h \int_0^t \nabla(p+w)(x,\tau)\,d\tau \\ &\quad - 2(p+w)_t \int_0^t h_r(x,\tau)\,d\tau - 2h_t \int_0^t (p+w)_r(x,\tau)\,d\tau \\ &\quad - 2(p+w)h_r + \frac{2}{r}\,h_t + 2\nabla h\nabla g - 2(p+w)_r h - 2h_t g'(r), \end{aligned} \tag{4.5}$$

and

$$L_2(h) = -2h_r h + 2\nabla h \int_0^t \nabla h(x,\tau)\,d\tau - 2h_t \int_0^t h_r(x,\tau)\,d\tau. \tag{4.6}$$

Although operators L_1, L_2 seem to be rather complicated, they will not give us any "troubles," simply because they do not contain second-order derivatives of h. Thus we will "suppress them by $L_0(h)$" using Theorem 3.1.

By (4.4) we obtain

$$L(p+w+h) + L(p+w) = 2L(p+w) + [L_0(h) + L_1(p+w,h) + L_2(h)].$$

Hence (4.3) leads to

$$\begin{aligned} J_\lambda(w+h) - J_\lambda(h) &= \int_G [L_0(h) + L_1(p+w,h) + L_2(h)] \\ &\quad \times [2L(p+w) + L_0(h) + L_1(p+w,h) + L_2(h)]e^{2\lambda\beta}\,dx\,dt. \end{aligned} \tag{4.7}$$

In order to calculate the Fréchet derivative $J'_\lambda(w)(h)$ we should single out the "linear part," with respect to h, of the right-hand side of (4.7). Hence

$$J'_\lambda(w)(h) = 2\int_G L(p+w)[L_0(h) + L_1(p+w,h)]e^{2\lambda\beta}\,dx\,dt. \tag{4.8}$$

Therefore (4.7) and (4.8) lead to

$$
\begin{aligned}
&J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \\
&\quad = \int_G [L_0(h)]^2 e^{2\lambda\beta}\, dx\, dt + \int_G L_0(h)[L_1(p+w,h) + L_2(h)] e^{2\lambda\beta}\, dx\, dt \\
&\quad + \int_G L_1(p+w,h)[L_0(h) + L_1(p+w,h) + L_2(h)] e^{2\lambda\beta}\, dx\, dt \\
&\quad + \int_G L_2(h)[2L(p+w) + L_0(h) + L_1(p+w,h) + L_2(h)] e^{2\lambda\beta}\, dx\, dt.
\end{aligned} \tag{4.9}
$$

Denote

$$
I_1(h,\lambda) = \int_G L_0(h)[L_1(p+w,h) + L_2(h)] e^{2\lambda\beta}\, dx\, dt, \tag{4.10}
$$

$$
I_2(h,\lambda) = \int_G L_1(p+w,h)[L_0(h) + L_1(p+w,h) + L_2(h)] e^{2\lambda\beta}\, dx\, dt, \tag{4.11}
$$

and

$$
I_3(h,\lambda) = \int_G L_2(h)[2L(p+w) + L_0(h) + L_1(p+w,h) + L_2(h)] e^{2\lambda\beta}\, dx\, dt. \tag{4.12}
$$

Hence (4.9) can be rewritten as

$$
\begin{aligned}
&J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \\
&\quad = \int_G [L_0(h)]^2 e^{2\lambda\beta}\, dx\, dt + I_1(h,\lambda) + I_2(h,\lambda) + I_3(h,\lambda).
\end{aligned} \tag{4.13}
$$

In order to prove the theorem we have to estimate the right-hand side of (4.13) from below. First, we estimate the integrals I_1, I_2, I_3 separately. As in §3, $C > 1$ denotes different positive constants dependent only on G, P'_N, and α. Let M be a positive constant such that

$$
\max\{1, \|w\|_{C^1(\bar{G})}, \|p+w\|_{C^2(\bar{G})}\} \le M, \qquad w \in P'_N. \tag{4.14}
$$

Formally speaking we could replace M with C, but we will not do that in order to make our estimates more clear. We remember that $h \in P'_N$. Hence

$$
\|h\|_{C^1(\bar{G})} \le M. \tag{4.15}
$$

First, we estimate from above the following functions:

$$
[L_1(p+w,h)]^2, \quad [L_2(h)]^2, \quad |L_2(h)|, \quad w, \quad h \in P'_N.
$$

By (4.5) and (4.14),

$$
\begin{aligned}
|L_1(p+w,h)| &\le 2M \int_0^t |\nabla h(x,\tau)|\, d\tau \\
&+ 2M \int_0^t |h_\tau(x,\tau)|\, d\tau + MC|h_t| + MC|\nabla h| + 2M|h|.
\end{aligned} \tag{4.16}
$$

Hence

$$|L_1(p+w,h)|^2 \le CM^2 \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left(\int_0^t |\nabla h(x,\tau)|\, d\tau \right)^2 \right\}. \tag{4.17}$$

Furthermore (4.6) and (4.15) imply

$$|L_2(h)| \le 2M \int_0^t |\nabla h(x,\tau)\, d\tau + 2M \int_0^t |h_r(x,\tau)|\, d\tau + 2M|h_r|.$$

Hence

$$|L_2(h)|^2 \le CM^2 \left\{ h_r^2 + \left(\int_0^t |\nabla h(x,\tau)|\, d\tau \right)^2 \right\}. \tag{4.18}$$

On the other hand, (4.6) and the Cauchy inequality imply

$$|L_2(h)| \le C \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left(\int_0^t (\nabla h(x,\tau))\, d\tau \right)^2 \right\}. \tag{4.19}$$

Now we are ready to estimate integrals I_1, I_2, and I_3. Using the Cauchy–Schwarz inequality and (4.10) we obtain

$$\begin{aligned} I_1(h,\lambda) &\ge -\frac{1}{4} \int_G |L_0 h|^2 e^{2\lambda\beta}\, dx\, dt \\ &\quad - \int_G [L_1(p+w,h) + L_2(h)]^2 e^{2\lambda\beta}\, dx\, dt \\ &\ge -\frac{1}{4} \int_G (L_0 h)^2 e^{2\lambda\beta}\, dx\, dt - 2 \int_G [L_1(p+w,h)]^2 e^{2\lambda\beta}\, dx\, dt \\ &\quad - 2 \int_G [L_2(h)]^2 e^{2\lambda\beta}\, dx\, dt. \end{aligned}$$

Furthermore, applying (4.17) and (4.18), we get

$$\begin{aligned} I_1(h,\lambda) &\ge -\frac{1}{4} \int_G (L_0 h)^2 e^{2\lambda\beta}\, dx\, dt \\ &\quad - CM^2 \int_G \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left[\int_0^t |\nabla h(x,\tau)|\, d\tau \right]^2 \right\} e^{2\lambda\beta}\, dx\, dt. \end{aligned} \tag{4.20}$$

Next, consider the integral I_2. The Cauchy–Schwarz inequality and (4.11) lead to

$$\begin{aligned} I_2(h,\lambda) &\ge -\frac{1}{4} \int_G (L_0 h)^2 e^{2\lambda\beta}\, dx\, dt - \int_G [L_1(p+w,h)]^2 e^{2\lambda\beta}\, dx\, dt \\ &\quad + \int_G [L_1(p+w,h)]^2 e^{2\lambda\beta}\, dx\, dt - \frac{1}{2} \int_G [L_1(p+w,h)]^2 e^{2\lambda\beta}\, dx\, dt \\ &\quad - \frac{1}{2} \int_G [L_2(h)]^2 e^{2\lambda\beta}\, dx\, dt \\ &= -\frac{1}{4} \int_G (L_0 h)^2 e^{2\lambda\beta}\, dx\, dt - \frac{1}{2} \int_G [L_1(p+w,h)]^2 e^{2\lambda\beta}\, dx\, dt \\ &\quad - \frac{1}{2} \int_G [L_2(h)]^2 e^{2\lambda\beta}\, dx\, dt. \end{aligned}$$

Hence applying inequalities (4.17) and (4.18), we obtain

$$
\begin{aligned}
I_2(h,\lambda) \geq & -\frac{1}{4}\int_G (L_0 h)^2 e^{2\lambda\beta}\,dx\,dt \\
& - CM^2 \int_G \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left(\int_0^t |\nabla h(x,\tau)|\,d\tau \right)^2 \right\} e^{2\lambda\beta}\,dx\,dt.
\end{aligned} \tag{4.21}
$$

Finally, we have to estimate the integral I_3. By (4.12),

$$
\begin{aligned}
I_3(h,\lambda) = & \int_G 2L(p+w)L_2(h)e^{2\lambda\beta}\,dx\,dt + \int_G L_0(h)L_2(h)e^{2\lambda\beta} \\
& + \int_G L_2(h)[L_1(p+w,h) + L_2(h)]e^{2\lambda\beta}\,dx\,dt.
\end{aligned} \tag{4.22}
$$

In order to estimate the first integral in (4.22) we apply (4.14) and (4.19). Thus

$$
\begin{aligned}
& 2\left| \int_G L(p+w)L_2(h)e^{2\lambda\beta}\,dx\,dt \right| \\
& \quad \leq CM \int_G |L_2(h)|e^{2\lambda\beta}\,dx\,dt \\
& \quad \leq CM \int_G \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left(\int_0^t |\nabla h(x,\tau)|\,d\tau \right)^2 \right\} e^{2\lambda\beta}\,dx\,dt.
\end{aligned} \tag{4.23}
$$

Hence the Cauchy–Schwarz inequality (4.17), (4.18), (4.22), and (4.23) lead to

$$
\begin{aligned}
I_3(H,\lambda) \geq & -\frac{1}{4}\int_G (L_0 h)^2 e^{2\lambda\beta}\,dx\,dt \\
& - CM^2 \int_G \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left(\int_0^t |\nabla h(x,\tau)\,d\tau \right)^2 \right\} e^{2\lambda\beta}\,dx\,dt.
\end{aligned}
$$

Thus estimates (4.20), (4.21), and the last inequality imply

$$
\begin{aligned}
& I_1(h,\lambda) + I_2(h,\lambda) + I_3(h,\lambda) \geq -\frac{3}{4}\int_G (L_0 h)^2 e^{2\lambda\beta}\,dx\,dt \\
& - CM^2 \int_G \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left(\int_0^t |\nabla h(x,\tau)|\,d\tau \right)^2 \right\} e^{2\lambda\beta}\,dx\,dt.
\end{aligned}
$$

Hence by (4.13) we obtain

$$
\begin{aligned}
& J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \geq \frac{1}{4}\int_G [L_0(h)]^2 e^{2\lambda\beta}\,dx\,dt \\
& - CM^2 \int_G \left\{ |\nabla h|^2 + h_t^2 + h^2 + \left(\int_0^t |\nabla h(x,\tau)|\,d\tau \right)^2 \right\} e^{2\lambda\beta}\,dx\,dt.
\end{aligned}
$$

Furthermore, applying inequality (3.32) and Lemma 3.8 we conclude from the last estimate that

$$
\begin{aligned}
& J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \\
& \quad \geq \frac{1}{4}\int_G (L_0 h)^2 e^{2\lambda\beta}\,dx\,dt - CM^2 \int_G (|\nabla h|^2 + h_t^2) e^{2\lambda\beta}\,dx\,dt.
\end{aligned} \tag{4.24}
$$

Now we are ready to apply the Carleman estimate from Theorem 3.1 to the first integral in (4.24). We obtain

$$\begin{aligned} &J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \\ &\quad \geq C\lambda \int_G (|\nabla h|^2 + h_t^2) e^{2\lambda\beta}\, dx\, dt - CM^2 \int_G (|\nabla h|^2 + h_t^2) e^{2\lambda\beta}\, dx\, dt \\ &\quad - Ce^{-2\lambda R} \int_{\omega_3} (|\nabla h|^2 + h_t^2)\, d\sigma \end{aligned}$$

for all $w, h \in P'_N$ and for all $\lambda \geq \hat{\lambda}$. Let $\lambda_0 = \max(\hat{\lambda}, 2M^2)$. Then the last inequality leads to

$$\begin{aligned} &J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \\ &\quad \geq \frac{1}{2} C\lambda \int_G (|\nabla h|^2 + h_t^2) e^{2\lambda\beta} - Ce^{-2\lambda R} \int_{\omega_3} (|\nabla h|^2 + h_t^2)\, d\sigma, \\ &\qquad w, \quad h \in P'_N, \quad \lambda \geq \lambda_0. \end{aligned} \tag{4.25}$$

Since for $v \in H^2(D_{T_1})$

$$\int_{\omega_3} (|\nabla v|^2 + v_t^2)\, d\sigma \leq C\|v\|_{H^2(D_{T_1})},$$

(4.25) implies

$$\begin{aligned} &J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \\ &\quad \geq \frac{1}{2} C\lambda \int_G (|\nabla h|^2 + h_t^2) e^{2\lambda\beta}\, dx\, dt - Ce^{-2\lambda R} \|h\|^2_{H^2(D_{T_1})}, \\ &\qquad w, \quad h \in P'_N, \quad \lambda \geq \lambda_0. \end{aligned} \tag{4.26}$$

Let $R_1 = R_1(N), R_1 \in (0, R)$ be the number defined in Lemma 2.1. Choose an arbitrary number $R_2 \in (R_1, R)$ and denote

$$G(R_2) = \left\{(x,t) : \beta(x,t) > -R_2, |x| > r_0, \sin\varphi > \frac{r_0}{R}, t > 0\right\}.$$

Since $\beta(x,t) = -(r + \alpha t)$ then

$$G(R_2) = \{(x,t) : -(r+\alpha t) > -R_2, |x| > r_0, t > 0\}.$$

Thus $G(R_2) \subset G$ and

$$\exp(2\lambda\beta) > \exp(-2\lambda R_2) > \exp(-2\lambda R) \quad \text{in } G(R_2). \tag{4.27}$$

Let $E = E(R_1, R_2)$ be the time cylinder

$$E = \left\{(x,t) : r_0 < |x| < R_1, \sin\varphi > \frac{r_0}{R}, 0 < t < \frac{R_2 - R_1}{\alpha}\right\}.$$

Hence $E \subset G(R_2)$ and by Lemma 2.3 there exists a positive constant $\gamma = \gamma(E, D_{T_1}, P_N)$ such that

$$\|f\|_{H^1(E)} \geq \gamma \|f\|_{H^2(D_{T_1})}, \qquad f \in P_N. \tag{4.28}$$

Note that since $R_1 = R_1(N)$ depends on P_N only, we can actually choose γ dependent only on D_{T_1} and P_N.

Therefore (4.26)–(4.28) lead to

$$\begin{aligned}
&J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \\
&\geq \frac{1}{2} C\lambda \int_{G(R_2)} (|\nabla h|^2 + h_t^2) e^{2\lambda\beta}\, dx\, dt - Ce^{-2\lambda R} \|h\|^2_{H^2(D_{T_1})} \\
&\geq \frac{1}{2} C\lambda \exp(-2\lambda R_2) \int_{G(R_2)} (|\nabla h|^2 + h_t^2)\, dx\, dt - Ce^{-2\lambda R} \|h\|^2_{H^2(D_{T_1})} \\
&\geq \frac{1}{2} C\lambda \exp(-2\lambda R_2) \int_{E} (|\nabla h|^2 + h_t^2)\, dx\, dt - Ce^{-2\lambda R} \|h\|^2_{H^2(D_{T_1})} \\
&\geq \frac{C\lambda\gamma}{2} \exp(-2\lambda R_2) \|h\|^2_{H^2(D_{T_1})} - Ce^{-2\lambda R} \|h\|^2_{H^2(D_{T_1})} \\
&= \frac{C\lambda\gamma}{2} \exp(-2\lambda R_2) \left[1 - \frac{2}{\lambda\gamma} \exp(-2\lambda(R - R_2))\right] \|h\|^2_{H^2(D_{T_1})}
\end{aligned}$$

for all $w, h \in P'_N$ and for all $\lambda \geq \lambda_0$. Since

$$\lim_{\lambda\to\infty} \left[\frac{2}{\lambda\gamma} \exp(-2\lambda(R - R_2))\right] = 0,$$

then we can choose $\tilde{\lambda} \geq \lambda_0$ such that

$$1 - \frac{2}{\lambda\gamma} \exp[-2\lambda(R - R_2)] > \frac{1}{2} \quad \text{for } \lambda \geq \tilde{\lambda}.$$

Therefore, we finally get the desired estimate

$$J_\lambda(w+h) - J_\lambda(w) - J'_\lambda(w)(h) \geq \frac{k}{2} \|h\|^2_{H^2(D_{T_1})}$$

for all $w, h \in P'_N$ and for all $\lambda \geq \tilde{\lambda}$, where

$$k = \frac{C\lambda\gamma}{2} \exp(-2\lambda R_2). \qquad \square$$

5. Global convergence of the gradient method on B_1. In this section we prove Theorems 1.2 and 1.3. Let us remind the reader that parameter $\lambda \geq \tilde{\lambda}$ is *fixed* in these theorems.

Proof of Theorem 1.2. Let $\{c_n\}$ be the sequence defined in (1.38). By the well-known formula

$$\begin{aligned}
\nabla F_\lambda(c_n) &= \nabla F_\lambda(c_*) + \int_0^1 H_\lambda(c_* + \tau(c_n - c_*))(c_n - c_*)\, d\tau \\
&= A_n(c_n - c_*),
\end{aligned} \tag{5.1}$$

the matrix A_n has the form

$$A_n = \int_0^1 H_\lambda(c_* + \tau(c_n - c_*))\, d\tau.$$

Hence (1.36) implies

$$k|y|^2 \le y^* A_n y \le K|y|^2, \qquad y \in \mathbb{R}^m. \tag{5.2}$$

Since

$$\nabla F_\lambda(c_*) = 0, \tag{5.3}$$

then (5.1) and (1.38) imply

$$\begin{aligned} |c_{n+1} - c_*| &= |c_n - c_* - \xi \nabla F_\lambda(c_n)| \\ &= |(I - \xi A_n)(c_n - c_*)| \le \|I - \xi A_n\| |c_n - c_*|, \end{aligned} \tag{5.4}$$

where I is identity matrix. Since $I - \xi A_n$ is a symmetric matrix, we have

$$\|I - \xi A_n\| = \max\{|1 - \mu_1|, |1 - \mu_m|\},$$

where μ_1 and μ_m are the smallest and the biggest eigenvalues of A_n, respectively. Hence (1.36), (1.39), and (1.40) lead to

$$\|I - \xi A_n\| \le q < 1. \tag{5.5}$$

Thus (5.4) and (5.5) imply

$$|c_{n+1} - c_*| \le q|c_n - c_*|. \tag{5.6}$$

Note that

$$|c_0 - c_*| = |0 - c_*| < \tfrac{1}{2}.$$

Hence by (5.5) and (5.6), $|c_n - c_*| < \frac{1}{2}$ for all n, which implies that $a_n \in B_1$ for all n. Finally, (5.6) leads to

$$|c_n - c_*| \le |0 - c_*| q^n < \tfrac{1}{2} q^n. \qquad \square$$

Proof of Theorem 1.3. In the case of *noisy* data we cannot claim that $\nabla F_\lambda(C_*) = 0$, where c_* is the minimum of J_λ on $\bar{B}_1$ with the *exact* data. However, by (4.8) the gradient $\Delta F_\lambda(c)$ is continuous on $\bar{B}_1$. Hence (1.22), (1.27), and (1.28) imply

$$|\nabla F_\lambda(c_*)| \le c_1 \delta, \tag{5.7}$$

where the positive constant C_1 depends only on D_{T_1}, P'_N, α, and λ. Choose a positive number $\hat{\delta}$ such that

$$\frac{C_1 \xi}{1 - q} \delta < \frac{1}{4} \quad \text{for all } \delta \in (0, \hat{\delta}]. \tag{5.8}$$

By (5.1) and (5.4) we obtain

$$\begin{aligned} |c_{n+1} - c_*| &= |c_n - c_* - \xi \nabla F_\lambda(c_n)| \\ &= |(I - \xi A_n)(c_n - c_*) - \xi \nabla F_\lambda(c_*)| \le \|I - \xi A_n\| |c_n - c_*| + \xi |\nabla F_\lambda(c_*)|. \end{aligned}$$

Hence (5.5) and (5.7) lead to

$$|c_{n+1} - c_*| \leq q|c_n - c_*| + C_1\xi\delta. \tag{5.9}$$

Furthermore, using the mathematical induction method and (5.9), we get

$$\begin{aligned} |c_n - c_*| &\leq |0 - a_*|q^n + C_1\xi\delta(1 + q + q^2 + \cdots + q^n) \\ &< \frac{1}{2}q^n + \frac{C_1\xi(1-q^n)}{1-q}\delta, \end{aligned}$$

as long as $a_n \subset B_1$. But (5.8) and the last inequality imply that $a_n \in B_1$ for all $n > 0$. Therefore

$$|c_n - c_*| < \frac{1}{2}q^n + C_1\frac{\xi(1-q^n)}{1-q}\delta$$

for all n. □

6. Discussion. The main result of this paper is Theorem 1.1, which claims that three-dimensional ISP (0.1), (0.2) with nonoverdetermined data can be reformulated as a locally convex optimization problem. We realize, however, some shortcomings of this result. The major shortcoming consists of imposing restrictive conditions on the wave field itself (function $v(x,t)$) rather than on the unknown coefficient $a(x)$. Generally speaking, we do not even know what these conditions imply for $a(x)$. Also, we cannot regard the "infinite tail" of Fourier series of the function $v(x,t)$ as a small noise because in this case one should assume that Fourier coefficients of $v(x,t)$ decay exponentially as $N \to \infty$ (due to the Carleman estimate). Note that if one would be able to handle the limit as $N \to \infty$, then one would get a global uniqueness result. We remember, however, that the global uniqueness result is a long term open problem for this ISP as well as for many similar three-dimensional problems with nonoverdetermined data. Once again, we wish to point out that Theorem 1.1 represents just a first result in this very difficult direction and this somehow justifies these shortcomings, at least in our opinions.

As to future efforts in this direction, one might try two lines of developments. The first one would consist of computational implementation of this or a similar numerical scheme. The second one would consist of relaxing stringent conditions of Theorem 1.1 by attempting to find a finite number of Fourier coefficients of the function $a(x)$ rather than of the wave field itself. Our preliminary studies show, however, that a more complicated numerical scheme should, perhaps, be considered in that case. Hence a combination of these two approaches rather than a single one should probably be implemented into computer codes.

Acknowledgment. We would like to express our gratitude to S. Gutman, M. Kazemi, R. Malek-Madani, and P. Sacks for fruitful discussions. Our special thanks to K. Kunisch and the Institute of Mathematics of Technical University of Graz, Austria, since the idea of this result occurred when M. Klibanov visited K. Kunisch at this department.

REFERENCES

[1] A. Baylis, V. Li, and C. S. Morawetz, *Scattering by potentials using hyperbolic methods*, Math. Comput., 52 (1989), pp. 321–328.

[2] A. L. Bukhgeim and M. V. Klibanov, *Uniqueness in the large of a class of multidimensional inverse problems*, Sov. Math. Dokl., 24 (1981), pp. 244–247.

[3] M. CHENEY, *A review of multi-dimensional inverse potential scattering*, in Inverse Problems in Partial Differential Equations, D. Colton, R. Ewing and W. Rundell, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[4] D. COLTON AND P. MONK, *The inverse scattering problem for acoustic waves in an inhomogeneous medium,* in Inverse Problems in Partial Differential Equations, D. Colton, R. Ewing, and W. Rundell, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[5] ———, *The numerical solution of an inverse scattering problem for acoustic waves*, IMA J. Appl. Math., 49 (1992), pp. 163–184.

[6] S. GUTMAN AND M. V. KLIBANOV, *Iterative method for multidimensional inverse scattering problems at fixed frequencies*, Inverse Problems, 10 (1994), pp. 573–599.

[7] M. R. HESTENES, *Conjugate Direction Method in Optimization*, Springer-Verlag, New York, 1980.

[8] L. HÖRMANDER, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1963.

[9] V. ISÀKOV, *Inverse Source Problems*, American Mathematical Society, Providence, RI, 1990.

[10] F. JOHN, *Continuous dependence on data for solutions of partial differential equations with a prescribed bound*, Comm. Pure Appl. Math., 13 (1960), pp. 551–585.

[11] M. KAZEMI AND M. V. KLIBANOV, *Stability estimates involving hyperbolic equations and inequalities*, Appl. Anal., to appear.

[12] M. V. KLIBANOV, *Inverse problems and Carleman estimates*, Inverse Problems, 8 (1992), pp. 575–596.

[13] M. V. KLIBANOV AND J. MALINSKY, *Newton-Kantorovich method for three-dimensional potential inverse scattering problem and stability of the hyperbolic Cauchy problem with time-dependent data,* Inverse Problems, 7 (1991), pp. 577–596.

[14] M. V. KLIBANOV AND RAKESH, *Numerical solution of a time-like Cauchy problem for the wave equation*, Math. Methods Appl. Sci., 15 (1992), pp. 559–570.

[15] M. V. KLIBANOV AND F. SANTOSA, *A computational quasi-reversibility method for Cauchy problems for Laplace's equation*, SIAM J. Appl. Math., 51 (1991), pp. 1653–1675.

[16] V. KOMORNIK AND E. ZUAZUA, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.

[17] I. LASIECKA AND R. TRIGGIANI, *Uniform stabilization of the wave equation with Dirichlet or Neuman feedback control without geometric conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.

[18] M. M. LAVRENT'EV, V. G. ROMANOV, AND S. P. SHISHATSKII, *Ill-Posed Problems of Mathematical Physics and Analysis*, American Mathematical Society, Providence, RI, 1986.

[19] V. G. ROMANOV, *Inverse Problems of Mathematical Physics*, VNU Press, Utrecht, the Netherlands, 1987.

[20] E. SOMERSALO, M. CHENEY, D. ISAAKSON, AND E. L. ISAAKSON, *A layer-stripping: a direct numerical method for impedance imaging*, Inverse Problems, 7 (1991), pp. 899–926.

[21] J. SYLVESTER, *A convergence layer stripping algorithm for the radially symmetric impedance tomography problem*, Comm. Partial Differential Equations, 17 (1992), pp. 1955–1994.

[22] W. SYMES, *Layered velocity inversion: a model problem from reflection seismology*, SIAM J. Math. Anal., 22 (1991), pp. 680–716.

[23] A. N. TIKHONOV AND V. YA. ARSENIN, *Solutions of Ill-Posed Problems*, Winston–Wiley, New York, 1977.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 180–189, January 1995

009

MULTIPLICITY RESULTS FOR TWO CLASSES OF BOUNDARY-VALUE PROBLEMS*

PHILIP KORMAN† AND TIANCHENG OUYANG‡

Abstract. Multiplicity results are provided for two classes of boundary-value problems with cubic nonlinearities, depending on a parameter λ. In particular, it is proved that for sufficiently large λ, there are exactly two solutions, and that all solutions lie on a single smooth solution curve. The last fact allows one to use continuation techniques to compute all solutions.

Key words. multiplicity results, bifurcation of solutions

AMS subject classification. 34B15

1. Introduction.

We consider a Dirichlet problem of the type

$$u'' + \lambda f(x,u) = 0 \quad \text{on } (a,b), \qquad u(a) = u(b) = 0 \tag{1}$$

for two classes of cubic nonlinearities depending on a parameter λ, and we prove existence and multiplicity results. We also study in detail the solution branches as $\lambda \to \infty$. For both types of nonlinearities we show existence of a critical λ_1, such that for $0 < \lambda < \lambda_1$, (1) has no nontrivial solution; it has at least one solution at $\lambda = \lambda_1$; and it has at least two solutions for $\lambda > \lambda_1$, with precisely two solutions for λ sufficiently large (nontrivial solutions that we find are positive by the maximum principle). Moreover, all solutions lie on a single curve of solutions. The last assertion is important for computational purposes, since it allows one to use efficient continuation techniques to compute all solutions of (1).

Exact multiplicity results are usually difficult to establish; see, e.g., Lions [5]. Our main tools are a bifurcation theorem of Crandall and Rabinowitz [2], and a variational argument due to Ambrosetti and Rabinowitz; see [7]. For both problems it is relatively easy to show that there are no solutions for sufficiently small $\lambda > 0$. We then show that for sufficiently large λ the functional corresponding to (1) has at least two critical points: a minimum point (corresponding to the stable maximal solution of (1)), and a saddle point (corresponding to the unstable minimum solution). To show that there are exactly two solutions for sufficiently large λ, we show that all solutions must lie on certain curves in the (λ, u) "plane." We then study the properties of these curves and exclude the possibility of more than two solutions.

The equations that we study have attracted considerable attention. For constant $a(x)$ and $b(x)$, problems (3) and (21) were studied by Smoller and Wasserman [10] (see also [11] and [12]), who obtained exact multiplicity results by a very nontrivial phase plane analysis. The Neumann problem for (3) was studied in detail by Angenent, Mallet-Paret, and Peletier [1] and Rocha [8]; see also Hale [3]. For f independent of x, both Neumann and Dirichlet problems were studied extensively by Schaaf [9].

Our approach appears to be quite general. We intend to consider other equations where exact multiplicity might be three or more for some values of λ. We are also working to extend our results to partial differential equations.

---

* Received by the editors June 1, 1992; accepted for publication (in revised form) April 28, 1993.

† Institute for Dynamics and Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221-0025.

‡ Department of Mathematics, Brigham Young University, Provo, Utah 84602.

Next we list some background results. Recall that a function $\varphi(x) \in C^2(a,b) \cap C^0[a,b]$ is called a supersolution of (1) if

$$\varphi'' + \lambda f(x,\varphi) \le 0 \quad \text{on } (a,b), \quad \varphi(a) \ge 0, \quad \varphi(b) \ge 0. \tag{2}$$

A subsolution $\psi(x)$ is defined by reversing the inequalities in (2). The following result is standard.

LEMMA 1. *Let $\varphi(x)$ and $\psi(x)$ be, respectively, super- and subsolutions of* (1), *and $\varphi(x) \ge \psi(x)$ on (a,b) with $\varphi(x) \not\equiv \psi(x)$; then $\varphi(x) > \psi(x)$ on (a,b).*

We shall often use this lemma with either $\varphi(x)$ or $\psi(x)$ or both being the solution of (1). The following lemma is a consequence of the first.

LEMMA 2. *Let $u(x)$ be a nontrivial solution of* (1) *with $f(x,0) \equiv 0$. If $u(x) \ge 0$ on (a,b) then $u > 0$ on (a,b).*

We proved the following proposition in [4].

PROPOSITION 1. *Consider the problem* (1) *and assume that $f(x,u) \in C^1([-1,1]\times R_+)$ satisfies*

(i) *$f(-x,u) = f(x,u)$ for all $x \in (-1,1)$ and $u > 0$;*

(ii) *$xf_x(x,u) < 0$ for all $x \in (-1,1)\backslash\{0\}$ and $u > 0$.*

Then any positive solution of (1) *is an even function with $u'(x) < 0$ on $(0, 1]$. Moreover, any two positive solutions of* (1) *do not intersect.*

Remark. Except for the last statement, this proposition is included in the Gidas–Ni–Nirenberg theorem.

Next we state a bifurcation theorem of Crandall and Rabinowitz [2].

THEOREM 1 [2]. *Let X and Y be Banach spaces. Let $(\bar\lambda, \bar x) \in \mathbf{R}\times X$ and let F be a continuously differentiable mapping of an open neighborhood of $(\bar\lambda, \bar x)$ into Y. Let the null-space $N(F_x(\bar\lambda, \bar x)) = \operatorname{span}\{x_0\}$ be one-dimensional and $\operatorname{codim} R(F_x(\bar\lambda, \bar x)) = 1$. Let $F_\lambda(\bar\lambda, \bar x) \notin R(F_x(\bar\lambda, \bar x))$. If Z is a complement of $\operatorname{span}\{x_0\}$ in X, then the solutions of $F(\lambda, x) = F(\bar\lambda, \bar x)$ near $(\bar\lambda, \bar x)$ form a curve $(\lambda(s), x(s)) = (\bar\lambda + \tau(s), \bar x + sx_0 + z(s))$, where $s \to (\tau(s), z(s)) \in \mathbf{R} \times Z$ is a continuously differentiable function near $s = 0$ and $\tau(0) = \tau'(0) = z(0) = z'(0) = 0$.*

Throughout this paper we consider only the classical solutions (which is not a serious restriction in the one-dimensional case). We also assume, without loss of generality, that $(a,b) = (-1,1)$.

2. A class of cubic nonlinearities with double root. On the interval $[-1,1]$ we consider the following boundary-value problem:

$$u'' + \lambda a(x)u^2(1 - b(x)u) = 0, \quad -1 < x < 1, \quad u(-1) = u(1) = 0. \tag{3}$$

We assume throughout this section that $a(x)$ and $b(x)$ are even functions $a(x) \in C^1(-1,1) \cap C^0[-1,1], b(x) \in C^2(-1,1) \cap C^0[-1,1]$, satisfying the following conditions:

$$a(x), b(x) > 0 \quad \text{for } -1 \le x \le 1; \tag{4}$$

$$xb'(x) > 0 \quad \text{and} \quad xa'(x) < 0 \quad \text{for } x \in (-1,1)\backslash\{0\}; \tag{5}$$

$$b''(x)b(x) - 2b'^2(x) > 0 \quad \text{for } -1 < x < 1. \tag{6}$$

For example, $b(x) = x^2 + \alpha$ with $\alpha > 3$ satisfies the above conditions. Notice that condition (6) implies that $1/b(x)$ is a supersolution of (3). To prove our multiplicity result we need the following lemmas. Recall that by maximum principle any solution of (3) is positive on $(-1,1)$.

LEMMA 3. *Every solution of* (3) *is strictly concave, i.e.,* $u'' < 0$ (*or* $1-b(x)u > 0$) *for all* $x \in (-1,1)$.

Proof. Denote $w(x) = b(x)u(x)$. Then one computes

$$w'' + \frac{\lambda a(x)}{b(x)} w^2(1-w) = 2b'u' + b''u. \tag{7}$$

If x_0 is a maximum point of $w(x)$, then

$$0 = w'(x_0) = b'(x_0)u(x_0) + b(x_0)u'(x_0),$$

i.e.,

$$u'(x_0) = -\frac{b'(x_0)}{b(x_0)} u(x_0).$$

Using this in (7), we obtain

$$w''(x_0) + \frac{\lambda a(x_0)}{b(x_0)} w^2(x_0)(1-w(x_0)) = \frac{u(x_0)}{b(x_0)}(b''(x_0)b(x_0) - 2b'^2(x_0)). \tag{8}$$

By our assumptions, the right-hand side of (8) is positive, while $w''(x_0) \le 0$. Hence $w(x_0) < 1$, i.e., $1 - b(x)u(x) > 0$ for all $x \in (-1,1)$, and the proof follows.

LEMMA 4. *Every solution of* (3) *is an even function with* $u'(x) < 0$ *for* $x \in (0,1]$.

Proof. Using Lemma 3 one sees that Proposition 1 applies, giving the conclusions of the lemma.

LEMMA 5. *Let* $u_\lambda(x)$ *be a continuous-in-*λ *branch of solutions of* (3). *Then either* $\lim_{\lambda\to\infty} u_\lambda(x) = 0$ *or* $\lim_{\lambda\to\infty} u_\lambda(x) = 1/b(x)$ *for all* $x \in (-1,1)$.

Proof. Rewrite (3) in the form

$$u_\lambda(x) = \lambda \int_{-1}^{1} G(x,\xi) a(\xi) u_\lambda^2(\xi)(1 - b(\xi)u_\lambda(\xi))\, d\xi, \tag{9}$$

where $G(x,\xi)$ is the corresponding Green's function, which is easily seen to be strictly positive and bounded on $(-1,1)\times(-1,1)$. By Lemma 3, $u_\lambda(x)$ is bounded as $\lambda \to \infty$ (by $1/b(x)$), and the integral on the right in (9) is positive. It follows that for each $\xi \in (-1,1)$ either $\lim_{\lambda\to\infty} u_\lambda(\xi) = 0$ or $\lim_{\lambda\to\infty} u_\lambda(\xi) = 1/b(\xi)$. Finally, since by Lemma 4 $u_\lambda'(\xi) < 0$ for $\xi \in (0,1)$, it follows that only one of the above possibilities holds for all ξ.

If $u(x)$ is a solution of (3), then the corresponding linearized problem will be used in the sequel

$$w'' + \lambda a(x)(2u - 3b(x)u^2)w = 0, \qquad w(-1) = w(1) = 0. \tag{10}$$

LEMMA 6. *If* (11) *has a nontrivial solution, then* $w(x)$ *does not change sign on* $(-1,1)$, *i.e., we can choose it so that* $w(x) > 0$ *on* $(-1,1)$.

Proof. Assume that $w(x)$ changes sign in $(-1,1)$. Assume that $w(x)$ has a zero on $[0,1)$, and the other case is similar. Without loss of generality (taking $-w$ if necessary), we may assume that $w(x) < 0$ on $(x_1,x_2), 0 \le x_1 < x_2 \le 1, w(x_1) = w(x_2) = 0$, and $w(x) > 0$ for $x < x_1$ and close to x_1, and for $x > x_2$ and close to x_2 (unless $x_2 = 1$). Differentiating (3), we obtain

$$(u')'' + \lambda a(x)(2u - 3b(x)u^2)u' = -\lambda a'u^2(1-bu) + \lambda ab'u^3. \tag{11}$$

Multiply (10) by u', (11) by w, and subtract and integrate both sides. Obtain

$$[w'u' - w(u')']|_{x_1}^{x_2} = \lambda \int_{x_1}^{x_2} [a'u^2(1-bu) - ab'u^3]w\,dx. \tag{12}$$

The quantity on the right side in (12) is positive by our assumptions. The one on the left is equal to

$$w'(x_2)u'(x_2) - w'(x_1)u'(x_1),$$

which is negative by Lemma 4. The contradiction proves the lemma.

LEMMA 7. *Let $u(x)$, the solution of* (3), *be such that* $\max_{[-1,1]} b(x)u(x) \leq \frac{1}{2}$. *Then the only solution of* (10) *is* $w \equiv 0$.

Proof. Since $u(x) > 0$ solves (3), it is the principal eigenfunction of

$$z'' + \lambda a(x)(u - b(x)u^2)z = \mu z, \qquad z(-1) = z(1) = 0,$$

corresponding to the principal eigenvalue $\mu = 0$. The principal eigenvalue of

$$w'' + \lambda a(x)(2u - 3b(x)u^2)w = \mu w, \qquad w(-1) = w(1) = 0 \tag{13}$$

must be positive, since $2u - 3bu^2 \geq u - bu^2$ for all $x \in (-1,1)$, with inequality being strict near $x = \pm 1$, by our assumption. If $w(x)$ is a nontrivial solution of (10), it is a nonprincipal eigenfunction of (13) (corresponding to $\mu = 0$), and so it must change sign on $[-1,1]$. But this contradicts the previous lemma.

THEOREM 2. *There exists a critical λ_1, such that for $0 < \lambda < \lambda_1$ the problem* (3) *has no solution; it has at least one solution at $\lambda = \lambda_1$; and it has at least two solutions for $\lambda > \lambda_1$. All solutions lie on a single curve of solutions, which is smooth in λ. For each $\lambda > \lambda_1$ there are finitely many solutions, and different solutions are strictly ordered on $(-1,1)$. Moreover, there exists $\lambda_2 \geq \lambda_1$, so that for $\lambda > \lambda_2$ the problem* (3) *has exactly two solutions denoted by $u^-(x,\lambda) < u^+(x,\lambda)$, with $u^+(x,\lambda)$ strictly monotone increasing in λ, $u^-(0,\lambda)$ strictly monotone decreasing in λ, and $\lim_{\lambda\to\infty} u^+(x,\lambda) = 1/b(x)$, $\lim_{\lambda\to\infty} u^-(x,\lambda) = 0$ for all $x \in (-1,1)$. (Recall that all solutions of* (3) *are positive by maximum principle.)*

Proof. Multiply (3) by u and integrate

$$\int_{-1}^{1} u'^2 dx = \lambda \int_{-1}^{1} a(x)u^2u(1-b(x)u)\,dx. \tag{14}$$

By the Poincaré inequality,

$$\int_{-1}^{1} u'^2 dx \geq \frac{\pi^2}{4} \int_{-1}^{1} u^2 dx.$$

On the other hand,

$$\int_{-1}^{1} a(x)u^2u(1-b(x)u)\,dx \leq \frac{a(0)}{4b(0)} \int_{-1}^{1} u^2\,dx.$$

Thus (3) has no solution for $\lambda < \pi^2 b(0)/a(0)$.

Existence of at least two solutions for sufficiently large λ follows similarly to the proof of a theorem of Ambrosetti and Rabinowitz; see [7, p. 12]. We outline the argument. Solutions of (3) are critical points on $H_0^1(-1,1)$ of the functional

$$J(u) = \int_{-1}^{1} \left(\frac{1}{2}u'^2 - \lambda a(x)\frac{u^3}{3} + \lambda a(x)b(x)\frac{u^4}{4} \right) dx.$$

It is easy to show that $J(u)$ is bounded from below, so that it must have a global minimum. By the Poincaré inequality, $J(u)$ is positive in a small neighborhood of zero in $H_0^1(-1,1)$. If we now can exhibit a function for which $J(u) < 0$, then in addition to a global minimum, where $J(u) < 0$, the functional $J(u)$ will have another critical point, where $J(u) > 0$, in view of the well-known mountain pass theorem; see [7]. It is easy to check that

$$J\left(\frac{1}{b(1)} \cos \frac{\pi}{2}x \right) < 0$$

for sufficiently large λ. (Alternatively, we could consider the evolution equation corresponding to (3) with the initial data

$$u(x,0) = \frac{1}{b(1)} \cos \frac{\pi}{2}x.$$

It is easy to show that $0 < u(x,t) \leq c$ for some $c > 0$, and so by well-known results, $u(x,t)$ would have to converge as $t \to \infty$ to the set of solutions of (3). Since $J(u(x,0)) < 0$ for sufficiently large λ, and $J(u(x,t))$ is nonincreasing in t, it follows that $u(x,t)$ cannot converge to zero. This would provide us with at least one positive solution of (3), which is sufficient for the arguments that follow.)

It is clear that the problem (3) has a maximal solution for λ large. We now study the curve of maximal solutions for decreasing λ. Rewrite (3) as

$$F(\lambda, u) = u'' + \lambda a(x)u^2(1 - b(x)u) = 0, \tag{15}$$

where $F : R \times C_0^2[-1,1] \to C[-1,1]$. Notice that $F_u(\lambda, u)w$ is given by the left-hand side of (10).

Now let $(\lambda_1, u(x))$ be a solution of (15). If the corresponding linearized equation (10) has only a trivial solution $w = 0$, then by the implicit function theorem we can solve (15) for $\lambda < \lambda_1$ and λ close to λ_1, obtaining a continuous-in-λ branch of solutions. We cannot continue this process of decreasing λ indefinitely, since we know that for $\lambda > 0$ sufficiently small, (15) has no solution. Let λ_0 be the infimum of λ for which we can continue the branch to the left. We claim there is a sequence $\{\lambda_n\}$ and $u_{\lambda_0} \in C_0^2(-1,1)$, a solution of (15) at $\lambda = \lambda_0$, so that as $\lambda_n \downarrow \lambda_0, u_{\lambda_n} \to u_{\lambda_0}$. Indeed, using Lemma 3, we conclude that there is a number $M > 0$, such that for any solution of (15),

$$\|u_\lambda\|_{C_0^2[-1,1]} \leq M.$$

It follows that a subsequence of $\{u_{\lambda_n}\}$ converges uniformly on $[-1,1]$. Passing to the limit in the integral version of (15) (see (9)), we establish the claim.

By the definition of λ_0 it follows that $F_u(\lambda_0, u_{\lambda_0})$ is singular, i.e., (10) has a nontrivial solution, which is positive by Lemma 6. By Lemma 6 one sees that $N(F_u(\lambda_0, u_{\lambda_0})) = \text{span}\{w(x)\}$ is one-dimensional, and then $\text{codim}R(F_u(\lambda_0, u_{\lambda_0})) = 1$,

since $F_u(\lambda_0, u_{\lambda_0})$ is a Fredholm operator of index zero. To apply the Crandall–Rabinowitz theorem (Theorem 1) it remains to check that $F_\lambda(\lambda_0, u_{\lambda_0}) \notin R(F_u(\lambda_0, u_{\lambda_0}))$. Assuming the contrary would imply the existence of $v(x) \not\equiv 0$, such that

$$v'' + \lambda_0(2au_0 - 3abu_0^2)v = au_0^2(1 - bu_0), \quad -1 < x < 1, \quad v(-1) = v(1) = 0. \tag{16}$$

Multiplying (16) by w, (10) by v, and integrating and subtracting, we obtain

$$\int_{-1}^{1} a(x)u_0^2(x)(1 - b(x)u_0(x))w(x)\,dx = 0,$$

which is a contradiction in view of Lemmas 3 and 6.

Applying Theorem 1, we conclude that $(\lambda_0, u_{\lambda_0})$ is a bifurcation point, near which the solutions of (3) form a curve $(\lambda_0 + \tau(s), u_{\lambda_0} + sw + z(s))$ with s near $s = 0$, and $\tau(0) = \tau'(0) = 0, z(0) = z'(0) = 0$. It follows that for λ close to λ_0 and $\lambda > \lambda_0$ we have two solutions $u^-(x, \lambda)$ and $u^+(x, \lambda)$ with $u^-(x, \lambda) < u^+(x, \lambda)$ for all $x \in (-1, 1)$, and that $u^+(x, \lambda)$ is strictly increasing in λ while $u^-(x, \lambda)$ is strictly decreasing. We show next that the upper branch $u^+(x, \lambda)$ is increasing in λ for all $\lambda > \lambda_0$. Differentiate (3) in λ:

$$u_\lambda'' + \lambda a(2u - 3bu^2)u_\lambda = -au^2(1 - bu), \qquad u_\lambda(-1) = u_\lambda(1) = 0. \tag{17}$$

We know by the above that $u_\lambda(x, \lambda) > 0$ for λ close to λ_0 and all $x \in (-1, 1)$. Let λ_1 be the first λ where this inequality is violated, i.e., $u_\lambda(x, \lambda_1) \geq 0$ and $u_\lambda(x_0, \lambda_1) = 0$ for some $x_0 \in (-1, 1)$. Applying the strong maximum principle to (17), we conclude that $u_\lambda(x, \lambda_1) > 0$ for all $x \in (-1, 1)$. Thus $u^+(x, \lambda)$ is strictly increasing in λ for all $\lambda > \lambda_0$.

After turning right the curve of solutions will decrease in λ, until a possible turn to the left occurs. If that happens, Theorem 1 applies exactly as above, and monotonicity of the branches follows similarly, so that after the turn the curve of solutions is increasing in λ (i.e., as we follow the curve for decreasing λ, the solution is decreasing). By the same reasoning as used previously, the curve will eventually have to turn to the right and decrease in λ, and so on. Denote by $(\lambda_i, u_i(x))$ the turning points (i.e., $F(\lambda_i, u_i)$ is singular).

We claim that the set of turning points is finite. Indeed, assuming the contrary, we first rule out a finite accumulation point $\bar{\lambda}$, i.e., $\lambda_{i_k} \to \bar{\lambda}$ along a subsequence. As previously, we show that a subsequence of u_{i_k} converges uniformly on $[-1, 1]$ to a solution $\bar{u}(x)$ of (3). Clearly $F_u(\bar{\lambda}, \bar{u})$ is singular (since otherwise the implicit function theorem would imply local uniqueness of the solution near $(\bar{\lambda}, \bar{u}(x))$). But then we have a contradiction with Theorem 1, which tells us that there can be no more than two solutions near $(\bar{\lambda}, \bar{u}(x))$. Next we rule out an infinite sequence of $\lambda_i \to \infty$. Notice that $u_{i+1}(x) < u_i(x)$ for all $i \geq 1$ and all $x \in (-1, 1)$. By Lemma 5, $u_i(x) \to 0$ as $i \to \infty$, but then we get a contradiction with Lemma 7, which tells us that there can be no bifurcations for sufficiently small u.

We now return to the curve of maximal solutions and follow it for increasing λ. If it turns to the left then Theorem 1 applies, and the curve is decreasing in λ after the turn (i.e., $u(x)$ is increasing when λ is decreasing). Since solutions of (3) are bounded, it follows as above that over any finite interval of λ's there is only a finite number of turns, and the final turn is to the right. Since all the while the solution is increasing, it follows by Lemma 5 that it approaches $1/b(x)$ as $\lambda \to \infty$. We show next that

for sufficiently large λ bifurcation is impossible, so that the curve of solutions keeps moving to the right in the (λ, u) "plane." Indeed, let $w(x)$ be a nontrivial solution of the linearized equation (10), normalized so that $\int_{-1}^{1} w^2\,dx = 1$. Multiply (10) by w, integrate by parts, and use the Poincaré inequality, obtaining

$$\int_{-1}^{1} a(x)(2u - 3b(x)u^2)w^2\,dx \geq \frac{\pi^2}{4\lambda}. \tag{18}$$

Since the quantity on the left is negative for u close to $1/b(x)$, we have a contradiction, which shows that (10) can have only trivial solution for λ large. (That $w(x)$ cannot concentrate near $x = +1$ follows similarly to Lemma 6.)

To recapitulate, we have a smooth curve of solutions which after a possible finite number of turns has a decreasing and single-valued-in-λ lower branch tending to zero, and a monotone increasing and single-valued-in-λ upper branch tending to $1/b(x)$. We show next that there is only one such curve. Indeed, assuming two such curves we would have for sufficiently large λ two upper branches, $v = v(x, \lambda)$ and $u = u(x, \lambda)$, both tending to $1/b(x)$. Denoting $w = u - v$, we express

$$w'' + p(x)w = 0 \quad -1 < x < 1, \quad w(-1) = w(1) = 0,$$

where $p(x) = a(x)[u + v - b(x)(u^2 + uv + v^2)]$ is negative for $u(x)$ and $v(x)$ close to $1/b(x)$. This leads to the same contradiction as previously, unless $w \equiv 0$.

Remark 2.1. Consider an interesting class of problems with the nonlinearity resembling the logistic one,

$$u'' + \lambda u^2(b(x) - u) = 0, \qquad u(-1) = u(1) = 0. \tag{19}$$

If $b(x)$ is an even function satisfying $b(x) > 0$ on $[-1, 1], b'(x) < 0$ for $x > 0$, and $b''(x) < 0$ for all $x \in (-1, 1)$, then it is easy to check that Theorem 2 applies.

Remark 2.2. Lemma 7 provides a lower estimate for the maximum value of any solution where bifurcation occurs, $u_m > 1/2b(0)$.

Remark 2.3. If u_m is the maximum value of the solution on the lower branch then

$$\frac{c_1}{\lambda} \leq u_m \leq \frac{c_2}{\lambda}, \quad \lambda > \lambda_2, \quad c_1, c_2 > 0. \tag{20}$$

Indeed, multiplying (3) by u and integrating,

$$\frac{\pi^2}{4}\int_{-1}^{1} u^2\,dx \leq \int_{-1}^{1} u'^2\,dx < \lambda a(0)u_m \int_{-1}^{1} u^2\,dx.$$

On the other hand, since all solutions are concave down, we have $u(x) \geq u_m|x - 1|$. Using this in (9), we easily obtain the second inequality in (20).

Remark 2.4. Based on the numerical evidence we believe that at $\lambda = \lambda_1$ the solution is unique, while for $\lambda > \lambda_1$ there are exactly two solutions.

3. Cubic nonlinearities with distinct roots. In this section we consider the problem

$$u'' + \lambda u(u - a(x))(b - u) = 0, \quad -1 < x < 1, \quad u(-1) = u(1) = 0. \tag{21}$$

Here b is a positive constant, and the function $a(x) \in C^1[-1,1]$ satisfies the following conditions:

$$(22)\quad a(x) \geq a_0 > 0, \quad a'(x) > 0 \quad \text{for } x \in (0,1), \quad a(-x) = a(x) \quad \text{for all } x \in (-1,1);$$

$$(23)\quad a(x) < \frac{1}{2}b \quad \text{for all } x \in (-1,1).$$

From the maximum principle every solution of (21) satisfies $0 < u < b$ in $(-1,1)$. Notice that, unlike (3), solutions of (21) are concave up near $x = \pm 1$.

LEMMA 8. *The solution of* (21) *is an even function. Moreover,* $u_x < 0$ *for* $x > 0$.

Proof. Since $0 < u(x) < b$ for all $x \in (-1,1)$, one easily sees that Proposition 1 applies.

LEMMA 9. *Let* $u(x,\lambda)$ *be a nontrivial solution of* (21) *for* $\lambda > \lambda_0$. *Then there are only three possibilities for* $\lim_{\lambda\to\infty} u(x,\lambda)$: $0, a(x)$, *and* b. *If the solution is increasing in* λ *then* $\lim_{\lambda\to\infty} u(x,\lambda) = b$ *for all* $x \in (-1,1)$.

Proof. The first part follows from the integral representation of the solution as before. From the previous lemma we know that for any $\lambda > \lambda_0, u(0,\lambda) > a(0)$. If the solution is increasing in λ this leaves us with $\lim_{\lambda\to\infty} u(0,\lambda) = b$. Indeed, the solution cannot tend to $a(x)$ over a subinterval, since $u_x < 0$ while $a'(x) > 0$, and it cannot tend to $a(x)$ at a point for the same reason.

As previously, we need to consider the linearization of (21),

$$(24)\quad w'' + \lambda[-3u^2 + 2(a+b)u - ab]w = 0, \quad -1 < x < 1, \quad w(-1) = w(1) = 0.$$

LEMMA 10. *If* (24) *has a nontrivial solution, we can choose it so that* $w(x) > 0$ *in* $(-1,1)$.

Proof. Assume on the contrary that $w(x)$ changes sign on $(-1,1)$. Assume $w(x)$ has a zero on $(-1,0]$ (the proof is similar if it has a root on $(0,1]$). We may then assume that $w(x) < 0$ on (x_1,x_2) with $-1 \leq x_1 < x_2 \leq 0$, and $w(x_1) = w(x_2) = 0, w'(x_1) < 0, w'(x_2) > 0$ (by changing if necessary to $-w$). Differentiate (21):

$$(25)\quad u_x'' + \lambda[-3u^2 + 2(a+b)u - ab]u_x = \lambda a'u(b-u).$$

Multiply (25) by w, (24) by u_x, and integrate and subtract:

$$(26)\quad (u_x'w - u_xw')|_{x_1}^{x_2} = \lambda \int_{x_1}^{x_2} a'(x)u(b-u)w\,dx.$$

The quantity on the right in (26) is positive by our assumptions, while the one on the left is

$$(27)\quad -u'(x_2)w'(x_2) + u'(x_1)w'(x_1) < 0$$

by Lemma 8.

THEOREM 3. *There exists a critical* λ_1, *such that for* $0 < \lambda < \lambda_1$ *the problem* (21) *has no solution; it has at least one solution at* $\lambda = \lambda_1$; *and it has at least two solutions for* $\lambda > \lambda_1$. *All solutions lie on a single smooth curve of solutions. For each* $\lambda > \lambda_1$ *there are finitely many solutions, and different solutions are strictly ordered. Moreover, there exists* $\lambda_2 \geq \lambda_1$ *so that for* $\lambda > \lambda_2$ *the problem* (21) *has exactly two solutions denoted by* $u^-(x,\lambda) < u^+(x,\lambda)$, *and* $\lim_{\lambda\to\infty} u^+(x,\lambda) = b$ *for all* $x \in (-1,1)$. *Solution* $u^-(x,\lambda)$ *develops a spike layer at* $x = 0$ *as* $\lambda \to \infty$.

Proof. The proof is similar to that of Theorem 2, so we shall not repeat all the details but concentrate on the points that are different. As before we show that (21) has no solutions for sufficiently small $\lambda > 0$. To show existence of at least two solutions for sufficiently large λ, we need to consider the functional

$$J(u) = \int_{-1}^{1} \left(\frac{1}{2}u'^2 + \lambda ab\frac{u^2}{2} - \lambda(a+b)\frac{u^3}{3} + \lambda\frac{u^4}{4} \right) dx$$

on $H_0^1(-1,1)$, and produce a function for which $J(u) < 0$. Consider the functional

$$\bar{J}(u) = \int_{-1}^{1} \left(ab\frac{u^2}{2} - (a+b)\frac{u^3}{3} + \frac{u^4}{4} \right) dx.$$

Using the condition (23) one computes $\bar{J}(b) < 0$. The function $u \equiv b$ does not satisfy the boundary conditions; however, it is clear that one can now construct $u_0(x) \in H_0^1(-1,1)$ with $\bar{J}(u_0(x))$ arbitrarily close to $\bar{J}(b)$, i.e., $\bar{J}(u_0) < 0$. Then for sufficiently large λ we have $J(u_0) < 0$, as desired.

To apply Theorem 1 it remains to verify that $F_\lambda(\lambda_0, u_{\lambda_0}) \notin R(F_u(\lambda_0, u_{\lambda_0}))$, where the map F and $(\lambda_0, u_{\lambda_0})$ are defined the same way as in the proof of Theorem 2. Assuming the contrary, we have $\int_{-1}^{1} u''w\,dx = 0$ (u is solution of (21), w of (24)). Notice that $w(x)$ is an even function (for otherwise the linear problem (24) would have another positive solution $w(-x)$, which is impossible). We then conclude that

$$\int_0^1 u''w\,dx = \int_0^1 u'w'\,dx = \int_0^1 uw''\,dx = 0.$$

Next we multiply (24) by xu_x, (25) by xw, and integrate and subtract. Using the above formula,

$$u'(1)w'(1) + \int_0^1 xa'(x)wu(b-u)\,dx = 0,$$

which is a contradiction, since both terms on the left are positive.

Proceeding as in the proof of Theorem 2, we follow the curve of maximal solutions left until a turning point $\lambda = \lambda_0$. Near that point, Theorem 1 implies existence of two solutions with $u^-(x,\lambda) < u^+(x,\lambda)$ for all $x \in (-1,1)$, and that u^- is decreasing in λ while u^+ is increasing in λ (for λ close to λ_0).

By Lemma 9, as $\lambda \to \infty$, any solution $u(x,\lambda)$ of (21) can only approach $0, b$, or $a(x)$. By Lemma 8, $u(x,\lambda)$ cannot approach $a(x)$ over any interval, since u_x and a' have opposite signs over $(-1,1)\backslash\{0\}$. On the other hand, $u(0,\lambda) > a(0)$, since $x = 0$ is the maximum point of $u(u_{xx}(0,\lambda) < 0)$. It follows that there are just two possibilities as $\lambda \to \infty$: either the solution approaches b for all $x \in (-1,1)$, or the solution approaches zero for $x \in (-1,1)\backslash\{0\}$, while $u(0,\lambda) > a(0)$, i.e., a spike-layer shape. (The possibility that $u^-(x,\lambda)$ approaches b on some proper subinterval of $(-1,1)$, and zero on its complement, is easily ruled out by the argument used in the proof of Proposition 1.)

As in Theorem 2 we show the existence of a smooth curve of solutions, which after possibly finitely many turns, has an upper branch $u^+(x,\lambda)$ single-valued in λ, and tending to b as $\lambda \to \infty$ (notice that for u close to b, (24) has only the trivial solution). The lower branch can also have only (possibly) finitely many turns, and it cannot tend to zero at a finite λ (as can be seen by converting (21) into an equivalent

integral equation). It is easy to see that the lower branch cannot approach b as $\lambda \to \infty$ (setting $w = u^+(x,\lambda) - u^-(x,\lambda)$, we obtain an equation similar to (24)). Hence the lower branch has to approach a spike layer shape described above. We next show that as this happens, further bifurcations (turns) are impossible. From (24) we obtain, as previously (normalizing w),

$$\int_{-1}^{1} [-3u^2 + 2(a+b)u - ab]w^2\,dx \geq \frac{\pi^2}{4\lambda}.$$

Since the quantity on the left is negative for u close to the spike layer, it follows that (24) has only the trivial solution.

We now have a smooth curve of solutions, which after a finite number of turns has an upper branch strictly monotone increasing and single-valued in λ and tending to b as $\lambda \to \infty$, and a lower branch single-valued in λ and tending to the spike-layer shape. We next show that there are no other solutions. Indeed, any other solution would have to lie on another curve of solutions, having the same properties. In particular, we would have another upper branch, tending to b, which was already ruled out previously.

Acknowledgment. It is a pleasure to thank Wei-Ming Ni for very useful discussions and encouragement.

REFERENCES

[1] S. B. ANGENENT, J. MALLET-PARET, AND L. A. PELETIER, *Stable transition layers in a semilinear boundary value problem*, J. Differential Equations, 67 (1987), pp. 212–242.

[2] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation, perturbation of simple eigenvalues and linearized stability*, Arch. Rational Mech. Anal., 52 (1973), pp. 161–180.

[3] J. K. HALE, *Asymptotic Behavior of Dissipative Systems*, AMS Mathematical Surveys and Monographs, No. 25, American Mathematical Society, Providence, RI, 1989.

[4] P. KORMAN AND T. OUYANG, *Exact multiplicity results for two classes of boundary problems*, Differential and Integral Equations, 6 (1993), pp. 1507–1517.

[5] P. L. LIONS, *On the existence of positive solutions of semilinear elliptic equations*, SIAM Rev., 24 (1982), pp. 441–467.

[6] T. OUYANG, *On the positive solutions of semilinear equations* $\Delta u + \lambda u + hu^p = 0$ *on compact manifolds. Part* I, Trans. Amer. Math. Soc., 331 (1992), pp. 503–527; *Part* II, Indiana Univ. Math. J., 40 (1991), pp. 1083–1141.

[7] P. H. RABINOWITZ, *Minimax Methods in Critical Point Theory with Applications to Differential Equations*, CBMS Reg. Conf. Ser. in Math. No. 65, American Mathematical Society, Providence, RI, 1986.

[8] C. ROCHA, *Examples of attractors in reaction-diffusion equations*, J. Differential Equations, 73 (1988), pp. 178–195.

[9] R. SCHAAF, *Global Solution Branches of Two Point Boundary Value Problems*, Lecture Notes in Math. 1458, Springer-Verlag, Berlin, 1990.

[10] J. SMOLLER AND A. WASSERMAN, *Global bifurcation of steady-state solutions*, J. Differential Equations, 39 (1981), pp. 269–290.

[11] S.-H. WANG AND N. D. KAZARINOFF, *Bifurcation and stability of positive solutions of two-point boundary value problem*, J. Austral. Math. Soc. Ser. A, 52 (1992), pp. 334–342.

[12] ———, *Bifurcation and steady-state solutions of a scalar reaction-diffusion equation in one space variable*, J. Austral. Math. Soc. Ser. A, 52 (1992), pp. 343–355.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 190–217, January 1995

010

BANANAS AND BANANA SPLITS: A PARAMETRIC DEGENERACY IN THE HOPF BIFURCATION FOR MAPS*

BRUCE B. PECKHAM†, CHRISTOS E. FROUZAKIS‡, AND IOANNIS G. KEVREKIDIS§

Abstract. The set of Hopf bifurcations for a two-parameter family of maps is typically a curve in the parameter plane. The side of the curve on which the invariant circle exists is further divided by horn-shaped resonance regions, with each region corresponding to maps that have a periodic orbit of a certain period. With the presence of a parametric degeneracy, the resonance regions sometimes take the form of closed "bananas" instead of open-ended horns. The authors investigate this local codimension-two bifurcation, emphasizing resonance regions as projections to the parameter plane of surfaces in phase × parameter space. The authors present scenarios where the degeneracy occurs "naturally" and illustrate them through an adaptive control application. More global implications of the local study are also discussed.

Key words. Hopf bifurcation, resonance, Arnold horns, parametric degeneracy

AMS subject classifications. 58F14, 58F40

1. Introduction.

When a fixed point for a map of $\mathbf{R}^n, n \geq 2$, has a complex conjugate pair of eigenvalues on the unit circle, we expect it to undergo a Hopf (also called Neimark–Sacker) bifurcation under perturbation. In a typical two-parameter family containing such a point, there is a Hopf bifurcation curve in the parameter plane which separates maps with an attracting fixed point from those with a repelling fixed point. The change in stability of the fixed point across the Hopf curve is accompanied, except possibly near the strong resonances, by the birth of an invariant topological circle from the fixed point. The side of the Hopf curve on which the invariant circle exists, as well as its stability, is determined by the relationship between the parameters, the linear terms, and some nonlinear terms in the family of maps. Again with the exception of parameter values near strong resonances, it is known that all local recurrence is restricted to the fixed point and to the invariant circle, when the latter exists.

On the side of the Hopf curve without the invariant circle, all nearby maps are locally topologically equivalent. On the side with the invariant curve, however, the parameter space must be further subdivided because, restricted to the invariant circle, we expect the rotation number of the maps, a topological invariant, to change with the parameters. From circle map theory we know that the existence of a reduced rational rotation number p/q implies the existence of at least one least-period-q orbit, so we concentrate in this paper on determining the location in phase × parameter space where periodic orbits of certain period exist. Such sets are called *period-q resonance surfaces*, or *p/q resonance surfaces* if we wish to identify the rotation number of the

* Received by the editors January 20, 1993; accepted for publication (in revised form) August 16, 1993.

† Department of Mathematics and Statistics, University of Minnesota at Duluth, Duluth, Minnesota 55812. This work was partially supported by National Science Foundation grant DMS 9020220.

‡ Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544. Present address: Institut für Energietechnik, Eidgenössische Technische Hochschule, Zentrum, CH-8092, Zürich, Switzerland. This work was partially supported by Defense Advanced Research Projects Agency, Office of Naval Research grant N00014-91-J-1850.

§ Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544. This work was partially supported by Defense Advanced Research Projects Agency, Office of Naval Research grant N00014-91-J-1850, and by a David and Lucile Packard Foundation Fellowship.

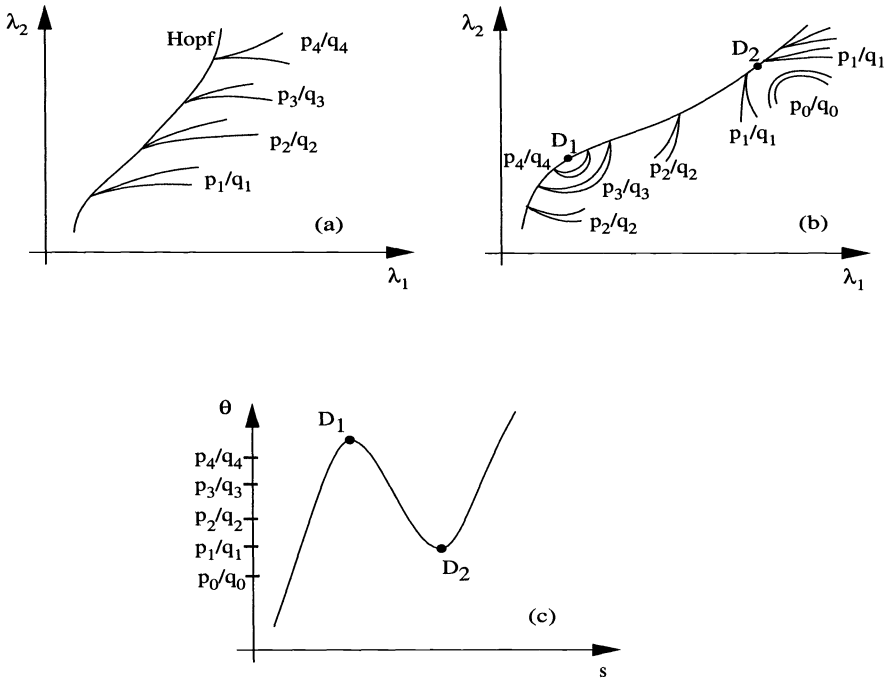


FIG. 1. *Typical Hopf bifurcations*: (a) *Parameter plane without angular degeneracies.* (b) *Parameter plane with two angular degeneracies (at D_1 and D_2).* (c) *Angle θ of fixed-point eigenvalue vs. arc length s along the Hopf curve in* (b).

period-q orbit. Their projections to parameter space are called *period-q resonance regions*, or *p/q resonance regions*, or (*Arnold*) *resonance horns* if we wish to suggest their shape. They are also called phase locking regions or entrainment regions, with both names originating in the context of Poincaré maps of two frequency flows on a torus; in this case, regions of constant frequency ratio for the flow correspond to regions of existence of a certain periodic orbit for the map.

It is known that a period-q resonance region typically opens out from every point on the Hopf curve for which the fixed point has an eigenvalue of $e^{2\pi ip/q}$, *with* p/q any reduced rational with $q \geq 5$, as suggested in Fig. 1a. In order to ensure that a specific resonance region opens in the horn–shaped manner suggested by that figure, several nondegeneracy conditions must be satisfied. Some nondegeneracy conditions pertain only to the phase variables, others include reference to the parameters as well. The failure of any one of the nondegeneracy conditions to hold results in a "degenerate" Hopf bifurcation. Of specific interest to us is the following nondegeneracy condition with respect to the parameters: it is usually assumed that the argument of the eigenvalues of the fixed point varies monotonically along the Hopf curve. When the argument *fails* to vary monotonically we say the Hopf bifurcation has an *angular degeneracy.* In this case, the nearby Arnold resonance regions can appear locally in shapes such as closed "bananas" rather than as the open horns of Fig. 1a. Figure 1b suggests a possible scenario for resonance regions near the Hopf bifurcation curve. The Hopf points with angular degeneracies are at D_1, a "banana" point, and at D_2, a "banana-split" point.

To emphasize the nonmonotonicity at the points with angular degeneracies, we show in Fig. 1c the argument of one of the eigenvalues of the neutral fixed point as a

function of arclength, s, along the Hopf bifurcation curve of Fig. 1b. The argument fails to vary monotonically through points D_1 and D_2.

To be more precise, we make the following definition. Unless otherwise noted, we assume throughout the paper that we are dealing with a k-parameter family of functions $\mathbf{F}_\mu : \mathbf{R}^2 \to \mathbf{R}^2, \mu \in \mathbf{R}^k$, which is C^∞ as a function from $\mathbf{R}^2 \times \mathbf{R}^k \to \mathbf{R}^2$. We will be mostly interested in two-parameter families ($k = 2$).

DEFINITION. Let $\mathbf{F}_\mu$ be a family of smooth maps of the plane with the following properties:

1. A map in the family has a fixed point

$$\mathbf{F}_{\mu_0}(\mathbf{x}_0) = \mathbf{x}_0.$$

2. The fixed point is nonhyperbolic, with complex conjugate eigenvalues on the unit circle. That is, the Jacobian

$$D\mathbf{F}_{\mu_0}(\mathbf{x}_0) \text{ has eigenvalues } e^{\pm 2\pi i \omega_0},$$

where $\omega_0 \in \mathbf{R}$, but $2\omega_0 \notin \mathbf{Z}$ to ensure that the eigenvalues in a neighborhood of the bifurcation point are complex.

Then $(\mathbf{x}_0, \mu_0)$ is a *Hopf bifurcation point* for the family $\mathbf{F}_\mu$.

The implicit function theorem guarantees that there exist unique fixed points near $\mathbf{x}_0$ for maps $\mathbf{F}_\mu$ corresponding to parameter values μ near μ_0. The fixed points can be described by a C^∞ function $\mathbf{x} = \mathbf{x}(\mu)$ satisfying $\mathbf{x}(\mu_0) = \mathbf{x}_0$. The eigenvalues of the nearby fixed point $\mathbf{x}(\mu)$ can be written as $\lambda_\pm = \lambda_\pm(\mu) = \lambda_\pm(\mathbf{x}(\mu)) = e^{\rho(\mu) \pm i(2\pi\omega_0 + \alpha(\mu))}$. This defines both $\rho(\mu)$ and $\alpha(\mu)$ uniquely, once a choice of ω_0 has been fixed, as C^∞ functions which must satisfy $\rho(\mu_0) = 0, \alpha(\mu_0) = 0$.

It is customary to study the Hopf bifurcation by making a change of parameters to (ρ, α) from the original parameters μ. This is possible whenever $\nabla_\mu \rho(\mu_0)$ and $\nabla_\mu \alpha(\mu_0)$ are linearly independent vectors.

DEFINITION. The point $(\mathbf{x}_0, \mu_0)$ is a *Hopf bifurcation point with a parametric degeneracy* if the vectors $\nabla_\mu \rho(\mu_0)$ and $\nabla_\mu \alpha(\mu_0)$ are *not* linearly independent.

DEFINITION. A Hopf bifurcation point satisfies the *eigenvalue crossing condition* if

$$\nabla_\mu \rho(\mu_0) \neq \mathbf{0}.$$

DEFINITION. We say $(\mathbf{x}_0, \mu_0)$ is a *bifurcation point with an angular degeneracy* for the family $\mathbf{F}_\mu$ if it has a parametric degeneracy, but the eigenvalue crossing condition is satisfied.

When the eigenvalue crossing condition is satisfied, as it generically is in two-parameter families, the implicit function theorem guarantees the continuation of a Hopf bifurcation curve through μ_0 in the parameter plane. If we express the Hopf curve with arc length parametrization as $\mu = \mu(s)$ with $\mu(0) = \mu_0$, it follows that $\rho(\mu(s)) = 0$.

If we monitor the argument of the neutral eigenvalue of the fixed point along the Hopf curve, we see that having an angular degeneracy is equivalent to $\frac{d}{ds}\alpha(\mu(s))|_{s=0} = 0$. This is why we call the degeneracy an *angular* degeneracy. An angular degeneracy occurs either when $\nabla_\mu \alpha(\mu_0)$ is nonzero but parallel to $\nabla_\mu \rho(\mu_0)$ or when $\nabla_\mu \alpha(\mu_0) = \mathbf{0}$ (the latter being a nongeneric occurrence in two-parameter families).

We think of the Hopf bifurcation with an angular degeneracy as arising from two possible scenarios in applications. The first is easier to explain and understand:

the "natural" parameters (μ above) in an application are *not* related in a one-to-one fashion to the "universal unfolding" parameters of the modulus and angle of the eigenvalues of the associated fixed point for the corresponding map, or equivalent parameters such as $\rho(\mu)$ and $\alpha(\mu)$ defined above. The lack of injectivity results in singular points for the change of parameters from μ to (ρ, α). Geometrically, we can think of curves of singular points as places we need to "fold" the natural parameter plane in order to place it on top of the corresponding points in the "universal" parameter plane. When the Hopf curve crosses such a fold curve in the natural parameter plane, we have an angular degeneracy. The description of the geometric "folding" of the parameter space is further detailed in §2.4.

The other general scenario where an angular degeneracy arises is along a curve of "secondary" Hopf bifurcations in the two-dimensional parameter space. Although locally the same as an angular degeneracy on a primary Hopf curve, this case cannot be dismissed as merely an "unfortunate" choice of parameters because a "good" parameter choice is usually determined with respect to primary bifurcation phenomena. One codimension-two bifurcation point, a Chenciner or transcritical Hopf point, *requires* the existence of an infinity of angular degeneracies, each on its own secondary Hopf curve inside its own resonance region. It was in studying bifurcations near a Chenciner point, in fact, when we first became interested in the angular degeneracy we describe in this paper [Mc], [Jo]. We discuss this scenario in more detail in §3. We also present in that section a model of a discrete-time adaptive control application that has a Chenciner point on a primary Hopf curve, a secondary Hopf curve connecting two Takens–Bogdanov points on the boundaries of a primary resonance region, an angular degeneracy on the secondary Hopf curve, and banana-shaped secondary resonance regions.

The main areas of emphasis of the paper are determining a model, or normal form, for a Hopf bifurcation with an angular degeneracy, investigating nearby resonance surfaces and their projections to parameter space, relating this bifurcation information for the model to the bifurcation picture for a generic two-parameter family of maps with a Hopf bifurcation with an angular degeneracy, and describing situations in which the angular degeneracy is expected to occur. The main result (Theorem 2.5 and its corollary) is that the resonance regions near a generic Hopf bifurcation point with an angular degeneracy "look" either like those near point D_1 or like those near point D_2 in Fig. 1b. We also discuss more global results about parameter space regions where banana resonance regions are expected to appear.

The paper is organized as follows. In §2, we recall some basic results about Hopf bifurcations, present the (known) Arnold theory for individual resonance horns (with emphasis on the resonance surfaces and using variations on Arnold's proofs), present analogous results for resonance regions near an angular degeneracy, and then consider the implications for the full bifurcation picture near an angular degeneracy. In §3, we describe scenarios in which secondary Hopf bifurcations with angular degeneracies are expected to occur, and present the adaptive control model. We discuss global parameter space bananas in §4, and make final comments in §5.

2. Local resonance regions near a Hopf bifurcation.

2.1. Background. We begin by recalling some standard terminology and results about Hopf bifurcations and normal forms.

DEFINITION. A *p/q resonant Hopf bifurcation point* is a Hopf bifurcation point having eigenvalues $e^{\pm 2\pi i p/q}$ and rotation number p/q around the fixed point for (an appropriate lift of) the linearization of the map at the Hopf point. The fraction p/q

must be in lowest terms. If $q \geq 5$, the bifurcation is said to be *weakly resonant*; if $3 \leq q \leq 4$, it is said to be *strongly resonant*. (Sometimes $q = 1$ and $q = 2$ are called strong resonances as well, although the eigenvalues for those cases are real.)

Note that, for a fixed choice of p/q, the p/q resonant Hopf bifurcation is a codimension-two bifurcation—one parameter is needed to bring the norm of the fixed-point eigenvalue to one, and the other parameter is needed to bring the argument of the fixed-point eigenvalue to the appropriate value of $2\pi p/q$.

THEOREM 2.1 (normal form theorem). *Let $(\mathbf{x}_0, \mu_0)$ be a Hopf bifurcation point for the k-parameter family of functions $\mathbf{F}_\mu : \mathbf{R}^2 \to \mathbf{R}^2$, C^∞ as a function from $\mathbf{R}^2 \times \mathbf{R}^k \to \mathbf{R}^2$. Then there exists a neighborhood of μ_0 in the parameter space for which the original family can be converted by a polynomial change of variables into the form*

$$\mathbf{f}_\mu(\mathbf{z}) = e^{\rho(\mu)+i\phi(\mu)}(\mathbf{z} + A(\mu)\mathbf{z}^2\overline{\mathbf{z}} + \cdots + B(\mu)\overline{\mathbf{z}}^{q-1} + \cdots), \tag{1}$$

by identifying $\mathbf{x} \in \mathbf{R}^2$ with $\mathbf{z} \in \mathbf{C}$, a translation (to bring the unique fixed point at each parameter value to the origin), and a "near identity" polynomial change of variables. $A(\mu) = A_1(\mu) + iA_2(\mu)$ and $B(\mu) = B_1(\mu) + iB_2(\mu)$ are complex valued functions. The omitted terms are all $O(|\mathbf{z}|^{q+1})$, except possibly those of the form $\mathbf{z}^j\overline{\mathbf{z}}^{j-1}, j \geq 3$ which are $O(|\mathbf{z}|^5)$. These "intermediate order" omitted terms are all invariant with respect to rotations; the $\overline{\mathbf{z}}^{q-1}$ term is the lowest-order term in the normal form which is not invariant with respect to all rotations. The dependence of all the functions with respect to μ is C^∞.

Proof. See [Ar], [GH], [Ru], for example. □

THEOREM 2.2 (Hopf bifurcation theorem). *Let $\mathbf{F}_\mu : \mathbf{R}^2 \to \mathbf{R}^2$ be a family of functions for which $\mathbf{F} : (\mu, \mathbf{x}) \to \mathbf{F}_\mu(\mathbf{x})$ is C^∞. Assume the eigenvalue crossing condition holds at $(\mathbf{x}_0, \mu_0)$ and that this point is not strongly resonant. Then*

1. *there is a unique fixed point near $\mathbf{x}_0$ for all maps near μ_0 in the parameter space. A C^∞-smooth Hopf bifurcation curve, defined by the neutral linear stability of the corresponding fixed point, passes through the point $(\mathbf{x}_0, \mu_0)$ in the parameter plane. The fixed point is stable on one side of the Hopf curve and unstable on the other side;*

2. *if in the normal form of equation (1) $A_1(\mu_0)$ is negative (positive), then an attracting (repelling) invariant circle surrounding the fixed point is born from the fixed point as the parameter crosses the Hopf bifurcation curve from the side with the attracting fixed point to the side with the repelling fixed point (from the side with the repelling fixed point to the side with the attracting fixed point). The smoothness of the invariant circles can be guaranteed to be C^r for any $r < \infty$ by suitably restricting the parameter space to a neighborhood of μ_0. Local recurrent points are the fixed point and some points on the invariant circle, when the circle exists.*

Proof. 1. The existence of a unique fixed point follows from the implicit function theorem. The stability follows from the Hartman–Grobman theorem.

2. See Ruelle's textbook [Ru] for a proof of this part of the theorem using the technique of graph transforms. □

When $A_1(\mu_0) < 0$, the Hopf bifurcation is called *supercritical*; when $A_1(\mu_0) > 0$, the bifurcation is called *subcritical*; when $A_1(\mu_0) = 0$, the bifurcation is called *transcritical* or a *Chenciner* point.

2.2. Individual nondegenerate resonance surfaces and regions (à la Arnold). The Hopf bifurcation theorem implies that the bifurcation study would be complete if we knew how to divide the parameter space on the side of the Hopf curve with the invariant circle into topological equivalence classes. Consequently, we

begin by studying surfaces in phase × parameter space corresponding to periodic orbits of a certain period. We first present a nondegenerate two-parameter model for investigating a "resonant" Hopf bifurcation in the neighborhood of a fixed point with eigenvalues $e^{\pm 2\pi ip/q}$. We will locate all local period-q points in a family containing such a resonant Hopf point; they will usually live on the invariant circle guaranteed by the Hopf bifurcation theorem. Although the results in this subsection are not new (cf. [Ar]), we include the subsection for several reasons: to emphasize the surfaces in phase × parameter space instead of just their projection to parameter space, to highlight the differences between the nondegenerate and the degenerate cases, to present some proofs which are slightly different from Arnold's proofs, and to make the paper more self-contained.

Arnold's analysis begins by studying vector fields which are invariant with respect to rotations of $e^{2\pi ip/q}$. He then shows that the qth iterates of maps such as the model families in (2) and (3) below are, up to arbitrarily high order, time-one maps of these equivariant vector fields. In contrast, we have chosen to work directly with maps, and for $q \geq 5$, although many of our arguments are suggested by his analysis, especially for his $q = 4$ case.

Our model family of maps near a p/q resonant Hopf bifurcation point is

$$\mathbf{f}_{(\rho,\alpha)}(\mathbf{z}) = e^{2\pi ip/q} e^{\rho + i\alpha}(\mathbf{z} + A\mathbf{z}^2\bar{\mathbf{z}} + B\bar{\mathbf{z}}^{q-1}), \tag{2}$$

where ρ and α are small real parameters, $\mathbf{z}$ is a complex variable, $\bar{\mathbf{z}}$ is its complex conjugate, p and q are integers, and $A = A_1 + iA_2$ and $B = B_1 + iB_2$ are complex constants with $A_1 \neq 0, B \neq 0$. We consider only the local bifurcation for $\mathbf{z}, \rho, \alpha$ near $\mathbf{0}, 0$, and 0, respectively. Our justification for using this model is in the proof of Corollary 2.4 below, where we show that a generic family near a p/q resonant Hopf point can be changed into the form of equation (2) plus some higher-order terms.

Properties of the (nondegenerate) resonant Hopf model. For the family defined by equation (2), which satisfies the hypotheses of the Hopf bifurcation theorem if $q \geq 5$, the fixed point $z = \mathbf{0}$ has eigenvalue $e^{\rho + i(\alpha + 2\pi p/q)}$. (The corresponding fixed point for the map in $\mathbf{R}^2$, obtained by identifying $\mathbf{R}^2$ with $\mathbf{C}$, has eigenvalues $e^{\rho \pm i(\alpha + 2\pi p/q)}$.) The line $\rho = 0$, where the origin $\mathbf{z} = 0$ changes from attracting ($\rho < 0$) to repelling ($\rho > 0$), is a Hopf bifurcation curve. The argument of the eigenvalue is monotonic along the Hopf curve. In fact, it equals $\alpha + 2\pi p/q$ at $(\rho, \alpha) = (0, \alpha)$. (Contrast this with the degenerate model, where this monotonicity fails to hold, in the next subsection.)

The local p/q resonance region, where period-q orbits exist, for equation (2) with $p/q = 1/5, A = -1 - i, B = 1$, is the horn-shaped region in Fig. 2. All three representative phase portraits are for the 5th iterate of the map. In phase portrait A, the 5th iterates rotate counterclockwise on the (attracting) invariant circle; in C they rotate clockwise; in B they move from saddles (×'s) toward nodes (filled circles).

More formally, we restate the following (known) theorem.

THEOREM 2.3. *Assume the family* $\mathbf{f}_{(\rho,\alpha)}$ *is defined as in* (2) *and* $q \geq 5$. *Then there exists a closed neighborhood* N *of the origin in the phase* × *parameter space with the following properties:*

1. *The set of least-period-*q *(*p/q*) points in* N *is topologically a punctured (closed) disk. The puncture point is the origin—the* p/q *resonant Hopf bifurcation point which is a fixed point. The union of the least-period-*q *points and the fixed point is a closed disk.*

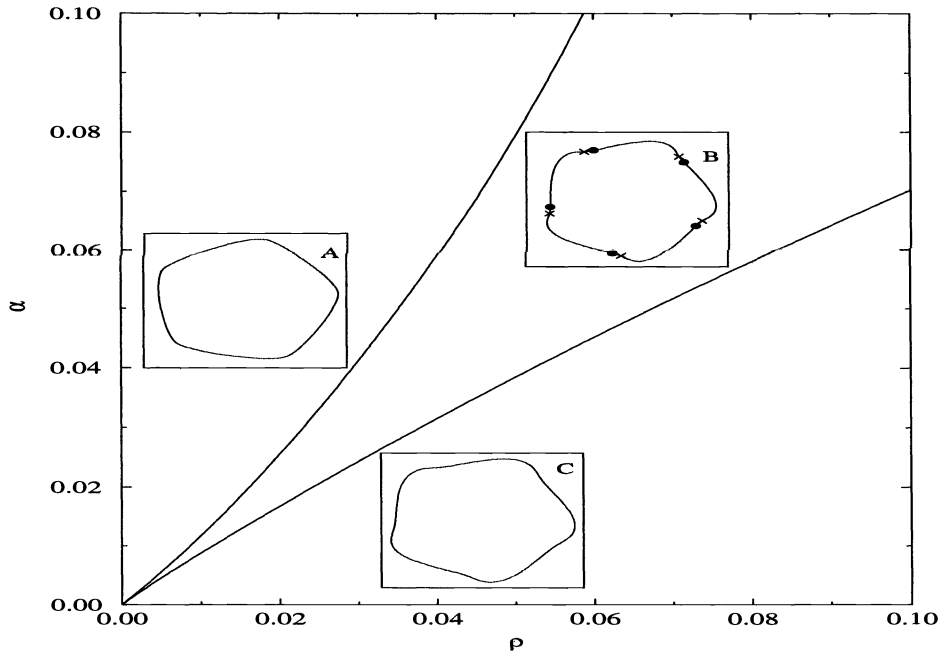


FIG. 2. *The nondegenerate Hopf model. Computations were done using* (2) *with* A=-1-i, B=1, *and* p/q=1/5.

2. *If* $A_1 \neq 0$, *then the projection of the least-period-q* (p/q) *points in* N *to the parameter space is an* (*Arnold*) *resonance horn, emanating from the origin in the parameter space, with both sides tangent to the vector* $(-A_1, -A_2)$. *If, in addition,* $B \neq 0$, *the horns have positive measure, and the order of tangency is* $\frac{q-2}{2}$. *The parameter values near* $(\rho, \alpha) = (0,0)$ *for which the corresponding maps have least-period-q orbit*(*s*) *near* $\mathbf{z} = \mathbf{0}$ *are precisely those inside and on the boundary of the resonance horn, excluding the tip of the horn, to which the* p/q *resonant Hopf point projects.*

3. *In the interior of this horn, there exists a pair of period-q* (p/q) *orbits, one attracting and one repelling when restricted to the invariant circle. The two orbits meet in a single saddle-node orbit on the boundaries of the horn* (*excluding the resonant Hopf point itself*).

Proof. 1. We determine all period-q points by looking at all solutions to $\mathbf{f}^q(\mathbf{z}) - \mathbf{z} = \mathbf{0}$ which are not fixed points. Expanding in terms of the parameters ρ and α and the modulus of $\mathbf{z}$, solving for the parameters, and neglecting higher-order terms leads to the result.

2. Eliminate the phase variables from the expressions obtained for the proof of the above item.

3. Follow arguments similar to those of Arnold [Ar].

Details are in the Appendix. □

COROLLARY 2.4. *Let* $\mathbf{F}_\mu : \mathbf{R}^2 \to \mathbf{R}^2$ *be a family of maps which satisfies the hypotheses of the Hopf bifurcation theorem as stated in* §2.1. *Assume also that* $(\mathbf{x}_0, \mu_0)$ *is a* p/q *weakly resonant Hopf bifurcation point* (*defined also in* §2.1), *without a parametric degeneracy* (*defined in the introduction*). *Then there is a neighborhood of* $(\mathbf{x}_0, \mu_0)$ *in the phase* × *parameter space in which the conclusions of Theorem* 2.3 *will hold,*

where in item 2, A is replaced by $A(\mu_0)$, B is replaced by $B(\mu_0)$, and the vector to which the resonance horn is tangent is the vector that $-A(\mu_0)$ is mapped to by the linearization of the coordinate change from the (ρ, α) parameter space to the original μ parameter space.

Proof. Change variables to bring the original equation into the form of equation (2) plus some higher-order terms. Details are in the Appendix. □

2.3. Resonance surfaces and regions near an angular degeneracy. Our model family of maps having least-period-q points near an angular degeneracy is

$$\mathbf{f}_{(\rho,\tau)}(\mathbf{z}) = e^{2\pi i\omega_0} e^{\rho + i(c_1\rho + c_2\tau^2)}(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + B\overline{\mathbf{z}}^{q-1}), \tag{3}$$

where ρ and τ are real parameters, $\mathbf{z}$ is a complex variable, $\overline{\mathbf{z}}$ is its complex conjugate, q is an integer, ω_0 is a real constant, c_1 and $c_2 \neq 0$ are real constants, and $A = A_1 + iA_2$ and $B = B_1 + iB_2$ are complex constants. As before, we consider only the local bifurcation for $\mathbf{z}, \rho, \tau$ near $\mathbf{0}, 0$, and 0, respectively. The use of this model is justified in Corollary 2.6.

As with the nondegenerate family in (2), this family has a Hopf bifurcation along $\rho = 0$. This family is "degenerate" because the argument of the fixed-point eigenvalue, $2\pi\omega_0 + c_1\rho + c_2\tau^2$, does not vary monotonically along the Hopf curve $\rho = 0$ as τ passes through zero. This causes a change in the appearance of the resonance regions, as we now describe in Theorem 2.5.

THEOREM 2.5 (properties of the degenerate Hopf model). *Assume the family $\mathbf{f}_{(\rho,\tau)}$ is defined as in (3) and $q \geq 5$. Assume A_1, c_2, and $A_2 - c_1A_1$ are all nonzero, and $\omega_0 \neq p/q$ but is sufficiently close to p/q. Then there exists a closed neighborhood N of the origin in the phase $\times$ parameter space with the following properties:*

1. *The set of least-period-q (p/q) points in N and the projection of this set to the parameter space are described by one of four cases. If we define $\alpha_0 := 2\pi(\omega_0 - p/q)$, then the four cases are determined by the signs of the three quantities $A_1, (A_2 - c_1A_1)/(c_2A_1)$, and $(\alpha_0A_1)/(A_2 - c_1A_1)$, as indicated respectively in the following list:*

   a. $(-,-,+)$ *or* $(+,+,-)$: *a twice-punctured sphere which projects to a banana-shaped region with both tips on the Hopf line.*

   b. $(+,-,+)$ *or* $(-,+,-)$: *two disjoint punctured closed disks, each projecting to disjoint resonance horns, each with its tip on the Hopf line (a "banana split").*

   c. $(-,+,+)$ *or* $(+,-,-)$: *a closed cylinder which projects to a "thickened" parabolic region.*

   d $(-,-,-)$ *or* $(+,+,+)$: *the empty set (projecting to the empty set).*

The punctures, present in the first two cases, are p/q resonant Hopf points located at $(\mathbf{z}, (\rho, \tau)) = (\mathbf{0}, (0, \pm\sqrt{-\alpha_0/c_2}))$, and project to corresponding horn tips. If $B \neq 0$, the parameter space horns have positive measure and have order of tangency $\frac{q-2}{2}$ at the tips. In all cases, the "centers" of the horns are pieces of parabolas to lowest order. The parameter values near $(\rho, \tau) = (0, 0)$, for which the corresponding maps have least-period-q (p/q) orbit(s) near $\mathbf{z} = \mathbf{0}$, are precisely those inside and on the boundary of the resonance region(s), excluding the resonant fixed points which project to the horn tips.

2. *On the interior of this region(s), there exists a pair of least-period-q (p/q) orbits, one attracting and one repelling, when restricted to the invariant circle. The two orbits meet in a single saddle-node orbit on the boundaries of the horn (excluding the resonant Hopf point(s) itself).*

Note: The terms in quotations are made more precise in the proof. Parameter space projections of cases (a), (b), and (c) are illustrated in Fig. 3a–3c, respectively.

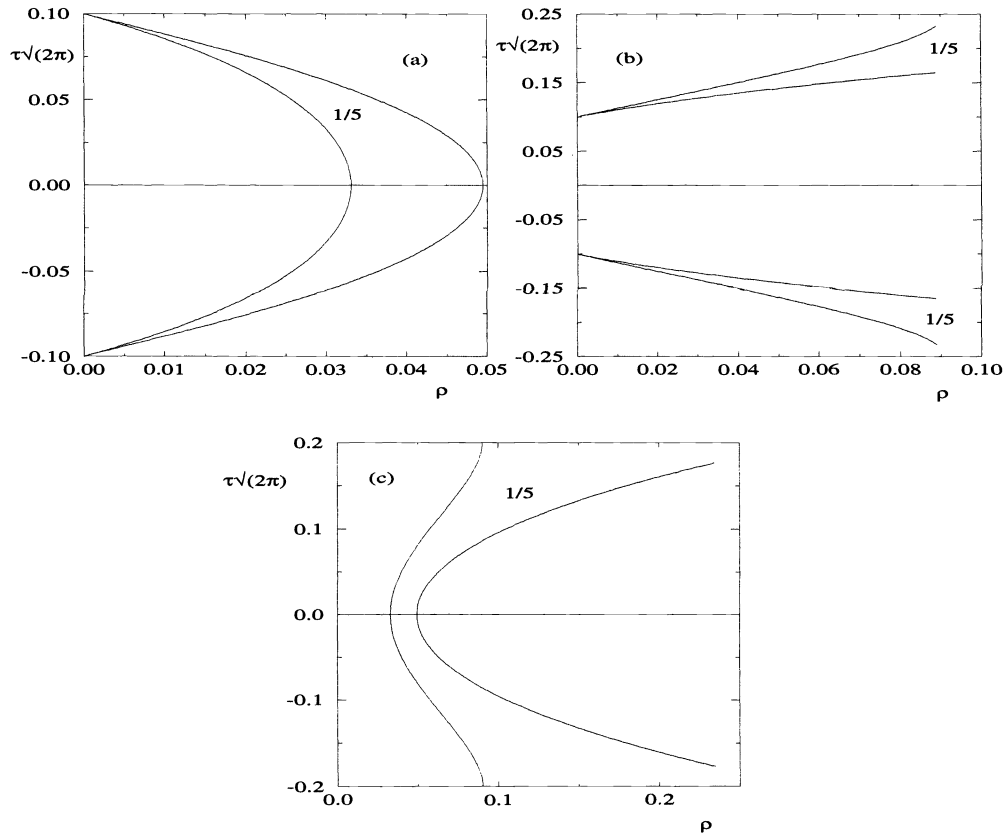


FIG. 3. *Period-5 resonance regions near a Hopf point with an angular degeneracy. Computations were done using equation* (3) *with* $A = -1 - i, B = 1, c_1 = -0.5$, *and* (a) $\omega_0 = 0.21, c_2 = -1$, (b) $\omega_0 = 0.19, c_2 = +1$, (c) $\omega_0 = 0.21, c_2 = +1$.

Proof. 1. The nondegenerate family of equation (2) and the degenerate family of equation (3) differ only in the appearance of their parameters: α has now been replaced by $\alpha_0 + c_1\rho + c_2\tau^2$. The proof is thus obtained by a (noninjective) change of parameters. Details are in the Appendix. See also the end of the next subsection, where with the aid of Fig. 4 we describe geometrically how the degenerate parameter space "unfolds" onto the nondegenerate parameter space.

2. Same as the proof of item 3 in Theorem 2.3. □

For the statement of the following corollary, we recall notation from the introduction: $\mu(s)$ is an arclength parametrization of the Hopf curve which passes through the bifurcation point at $s = 0$, and $e^{\rho(\mu)\pm i(2\pi\omega_o+\alpha(\mu))}$ are the eigenvalues of the corresponding fixed point along the Hopf curve.

COROLLARY 2.6. *Let* $\mathbf{F}_\mu : \mathbf{R}^2 \to \mathbf{R}^2$ *be a family of maps which satisfies the hypotheses of the Hopf bifurcation theorem as stated in* §2.1, *including the eigenvalue crossing condition. Assume that* $(\mathbf{x}_0, \mu_0)$ *is a Hopf bifurcation point with an angular degeneracy* (*defined in the introduction*), *the eigenvalues of* $D\mathbf{F}_{\mu_0}(\mathbf{x}_0)$ *are* $e^{\pm 2\pi i\omega_0}$, *the rotation number around* $\mathbf{x}_0$ *of* (*a lift of*) $D\mathbf{F}_{\mu_0}(\mathbf{x}_0)$ *is* $+\omega_0$, *and* $\omega_0 \neq p/q$ *for* $q \leq 4$ (*not strongly resonant*).

Assume also that and $\nabla_\mu\alpha(\mu_0) \neq \mathbf{0}$. *Thus* $\nabla_\mu\rho(\mu_0)$ *and* $\nabla_\mu\alpha(\mu_0)$ *are nonzero parallel vectors and* $\frac{d}{ds}\alpha(\mu(s))|_{s=0} = 0$. *Assume, however, that* $\frac{d^2}{ds^2}\alpha(\mu(s))|_{s=0} \neq 0$.

Then for any p/q sufficiently close to ω_0, there is a closed neighborhood N of $(\mathbf{x}_0, \mu_0)$ in the phase $\times$ parameter space inside which the set of least-period-q (p/q) points in N is described by one of the four cases (a)–(d) *enumerated in statement* 1 *of Theorem* 2.5. *The four cases are determined in the same way as in Theorem* 2.5, *after we put the original equation in its normal form up to $O(|z|^3)$ terms, change parameters from μ to (ρ, τ), let $A = A(\mu_0)$, and write the eigenvalue argument $\phi(\rho, \tau) = \omega_0 + c_1\rho + c_2\tau^2 + \cdots$, a form justified in the proof.*

Proof. This proof is similar to that of Corollary 2.4. We show that there is a nonsingular change of coordinates which brings our original equation into the same form as our model with an angular degeneracy except for higher-order terms. Details are in the Appendix. □

2.4. Discussion. Although the theorems and their corollaries in the previous subsection stated results for only one p/q resonance surface at a time, there are some relationships between nearby resonance surfaces we wish to point out.

First, for a p/q resonance horn away from an angular degeneracy, we recall from Corollary 2.4 that the angle at which it meets the Hopf curve is determined by the coefficient $A(\mu)$ of the $z^2\overline{z}$ term in the normal form, where μ is the parameter value corresponding to a p/q resonant Hopf point. Since this coefficient varies smoothly along the Hopf curve, the angles at which the various resonance horns meet the Hopf curve will also vary smoothly along the Hopf curve. (No similar statement can be made about the "$B(\mu)$" coefficient of the $\overline{z}^{q-1}$ term; it is not even the coefficient of the same term in the normal form as we move from one resonant Hopf point to another.) This implies that along the Hopf curve for any family, the angle at which a p/q resonance horn meets the Hopf curve varies smoothly as p/q varies. This is even true if the B coefficient in the normal form near a particular p/q resonant Hopf point is zero; the order of tangency of the saddle-node curves for that particular p/q resonance horn, however, would not be of order $\frac{q-2}{2}$.

A similar statement holds for the consistency in the shape of resonance regions near an angular degeneracy. The "parabolas" which define the "centers" of the p/q resonance regions (defined in the proof of Theorem 2.3) vary smoothly in p/q.

We also point out that, even though there are four distinct cases for individual resonance regions, the collection of resonance surfaces and regions near a single Hopf bifurcation with an angular degeneracy has one of the following two forms:

a. Twice-punctured disks which project to "bananas" for all p/q on one side of ω_0; the empty set for all p/q on the other side of ω_0.

b. Pairs of punctured disks which project to "banana splits" for all p/q on one side of ω_0; closed cylinders which project to thickened parabolas for all p/q on the other side of ω_0.

Analytically, this is because the signs of the three quantities A_1, $(A_2 - c_1A_1)/(c_2A_1)$, and $(\alpha_0A_1)/(A_2 - c_1A_1)$ determine the four cases; only the last quantity can change sign as p/q is varied (via $\alpha_0 := 2\pi(\omega_0 - p/q)$).

The first case is illustrated schematically in Fig. 1b near point D_1 and for the model family below in Fig. $4d_1$. The second case is illustrated schematically in Fig. 1b near point D_2 and for the model family below in Fig. $4d_2$. See also Fig. 10, described in the proof of Theorem 2.5 in the Appendix, for a further description of how nearby resonance regions change as varying p/q causes α_0 to change between positive and negative.

Model family. To portray a bifurcation picture with more than one resonance region near an angular degeneracy, we used the following family:

$$\mathbf{f}_{(\rho,\tau)}(\mathbf{z}) = e^{2\pi i\omega_0} e^{\rho+i(c_1\rho+c_2\tau^2)}(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + B\overline{\mathbf{z}}^{q-1} + C\mathbf{z}^3\overline{\mathbf{z}}) \tag{4}$$

with $\omega_0 = 0.19, c_1 = -0.5, A = -1 - i, B = 1, C = 1, q = 5$. Figure $4d_1$, using $c_2 = -1$, shows two banana resonance regions; Fig. $4d_2$, using $c_2 = +1$, shows a banana split resonance region and two parabolic resonance regions.

The family of equation (4) is the same as the model degenerate family we began with in equation (3), except for the $\mathbf{z}^3\overline{\mathbf{z}}$ term. We made this alteration because the family of equation (3) is invariant to rotations by $2\pi p/q$. This is fine for computing the p/q resonance region, but not for any other resonance region. For example, if $p/q = 1/5$, and we were computing the $1/6$ resonance region, the invariance with respect to rotations by $2\pi/5$ would imply that period-6 orbits must appear in groups of 5. Thus, a saddle-node birth of a pair of period-6 orbits would result in the birth of 10 period-6 orbits, or 60 period-6 points. The $\mathbf{z}^3\overline{\mathbf{z}}$ term was chosen because it is of high enough order so as not to affect the existence of the invariant circle, and because it is not invariant to any rotations about the origin in phase space. Thus no unwanted symmetries are present.

The geometry of the parameter change, or "Theorem 2.3 to Theorem 2.5 in pictures." Figures $4d_1$ and $4d_2$ can be thought of as having been created via parameter space "surgeries" of a bifurcation diagram for a corresponding nondegenerate family. Specifically, if we start with the bifurcation diagram of Fig. 4a for the nondegenerate family $\mathbf{f}_{(\rho,\alpha)}(\mathbf{z}) = e^{\rho+i\alpha}(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + B\overline{\mathbf{z}}^{q-1} + C\mathbf{z}^3\overline{\mathbf{z}})$, we can change it into either Fig. $4d_1$ or $4d_2$ with the coordinate change $\alpha = 2\pi\omega_0 + c_1\rho + c_2\tau^2$. This coordinate change, replacing α with τ, can be decomposed into the following three coordinate changes, each having a simple geometric interpretation:

a. Shear to make the "singular line" perpendicular to the Hopf curve: $\alpha = \hat{\alpha} + c_1\rho$ (Figs. 4a and 4b).

b. "Unfold a double cover of half of the nondegenerate parameter space": $\hat{\alpha} = 2\pi\omega_0 + c_2|\hat{\tau}|, c_2 = \pm 1$ (from Fig. 4b to $4c_1$ for $c_2 = 1$; from Fig. 4b to $4c_2$ for $c_2 = -1$).

c. Smooth the fold lines: $\hat{\tau} = \tau|\tau|$ (from Fig. $4c_1$ to $4d_1$, or from Fig. $4c_2$ to $4d_2$).

A rescaling would give the same picture as Fig. $4d_1$ for any negative c_2, and the same picture as Fig. $4d_2$ for any positive c_2.

It is now easier to see why the nondegeneracy conditions of Theorem 2.5 are necessary. The value of A_1 must be nonzero so that the resonance horns emerge transverse to the Hopf curve. The expression $A_2 - c_1A_1$ must be nonzero to ensure the resonance horns cross the fold line transversely (the horns emerge with slope $\frac{A_2}{A_1}$ and the slope of the fold line is c_1). If c_2 were zero, the fold might be even more degenerate.

More general degenerate families could also be considered as geometric unfoldings of a double cover of half a nondegenerate parameter space, but only to lowest-order terms. The model families behave better because they have constant coefficients A and B; in a more general family these coefficients would depend on the parameters.

3. Angular degeneracies on secondary Hopf bifurcation curves. So far, the only reason we have given for expecting a Hopf bifurcation with an angular degeneracy is that the relationship between an application's natural parameters and the "universal" parameters, the modulus and argument of a fixed point's eigenvalue,

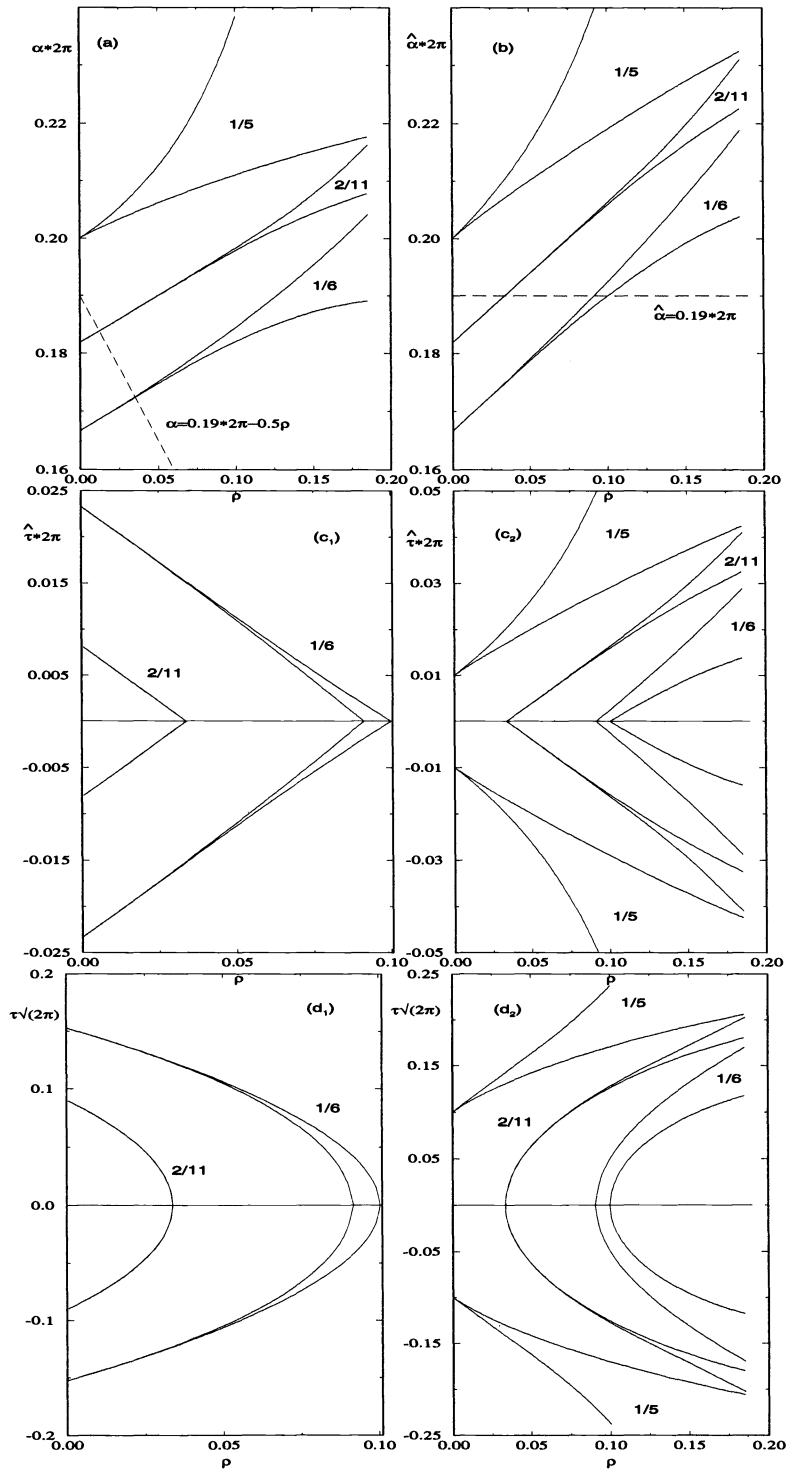


FIG. 4. *Relationship of the nondegenerate parameter space to the degenerate one*: (a) $\mathbf{f}_{(\rho,\alpha)}(\mathbf{z}) = e^{2\pi i\omega_0}e^{\rho+i\alpha}(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + B\overline{\mathbf{z}}^{q-1} + C\mathbf{z}^3\overline{\mathbf{z}})$; (b) $\mathbf{f}_{(\rho,\hat{\alpha})}(\mathbf{z}) = e^{2\pi i\omega_0}e^{\rho+i(\hat{\alpha}+c_1\rho)}(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + B\overline{\mathbf{z}}^{q-1} + C\mathbf{z}^3\overline{\mathbf{z}})$; (c) $\mathbf{f}_{(\rho,\hat{\tau})}(\mathbf{z}) = e^{2\pi i\omega_0}e^{\rho+i(2\pi\omega_0+c_1\rho+c_2|\hat{\tau}|)}(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + B\overline{\mathbf{z}}^{q-1} + C\mathbf{z}^3\overline{\mathbf{z}})$; (d) $\mathbf{f}_{(\rho,\tau)}(\mathbf{z}) = e^{2\pi i\omega_0}e^{\rho+i(c_1\rho+c_2\tau^2)}(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + B\overline{\mathbf{z}}^{q-1} + C\mathbf{z}^3\overline{\mathbf{z}})$ *In all figures,* $\omega_0 = 0.19, c_1 = -0.5, A = -1 - i, B = 1, C = 1, q = 5$; *in* c_1 *and* d_1, *the constant* $c_2 = -1$; *in* c_2 *and* d_2, *the constant* $c_2 = +1$.

could be nonhomeomorphic. We now describe some scenarios in which the angular degeneracy is expected, or even guaranteed, to occur, even for the "best" choice of parametrizations. They all involve secondary, rather than primary, Hopf bifurcations. These scenarios, in fact, were the original motivation behind our study of a Hopf bifurcation with an angular degeneracy which led to this paper.

3.1. Takens–Bogdanov points and secondary bifurcations. When a fixed point of a family $\mathbf{F}_\mu$ undergoes a (primary) Hopf bifurcation, one result can be the birth of periodic orbits as the Hopf curve in the parameter plane is crossed. Period-q resonance regions (horns), described throughout this paper, where period-q orbits exist, emanate from a point on the primary Hopf curve where the eigenvalues of $D\mathbf{F}_\mu$ at the associated fixed point are located at a qth root of unity.

The sides of a period-q resonance region are period-q saddle-node bifurcation curves, characterized by having an eigenvalue of $D\mathbf{F}_\mu^q$ at one. As a saddle-node curve is traced out in the parameter space, away from the primary Hopf bifurcation, the second eigenvalue may vary. (For $q \geq 5$ and parameter values near the primary Hopf bifurcation, the second eigenvalue determines the local attraction or repulsion normal to the invariant curve.) If the second eigenvalue also becomes equal to one, we generically have a double 1, or "Takens–Bogdanov" point [Bo], [Ta]. One consequence of the analysis of a generic Takens–Bogdanov point is the emergence of a (secondary) Hopf bifurcation curve from the Takens–Bogdanov point, tangent to the saddle-node and extending *into* the primary resonance region. This secondary Hopf bifurcation curve is characterized by the existence of a period-q point where the eigenvalues of $D\mathbf{F}_\mu^q$ are complex conjugate and on the unit circle. A (secondary) period-mq resonance horn, analogous to a primary period-m resonance horn, will emanate from a point on the secondary Hopf curve where $D\mathbf{F}_\mu^q$ at the associated period-q point has an eigenvalue at an mth root of unity.

Only five possibilities exist for the global continuation of a Hopf curve in a two-parameter family: (1) continuation in each direction terminates at a Takens–Bogdanov point; (2) continuation in each direction terminates at a "double (-1) point" [Ar], [Ta]; (3) continuation in one direction terminates at a Takens–Bogdanov point, continuation in the other direction terminates at a double (-1) point; (4) continuation forms a closed curve; or (5) continuation proceeds forever (in an unbounded parameter space). Possibilities (1) and (2) imply the existence of local extrema for the argument of the neutral eigenvalue along the Hopf curve. These local extrema are generically the Hopf bifurcation points with angular degeneracies.

Several possible scenarios, all involving secondary Hopf bifurcations and most involving Takens–Bogdanov points, are suggested in Figs. 5a–e. In Figs. 5a–d, we can assume the eigenvalue argument is zero at one of the Takens–Bogdanov points. Continuity of this eigenvalue along the secondary Hopf bifurcation curve, coupled with the assumed fact that no double (-1) points are encountered along the way, implies that the argument must return to zero at the other Takens–Bogdanov point. Thus the argument (generically nonconstant) must reach a maximum or minimum at least once along the secondary Hopf curve. In Fig. 5e, if by moving all the way around the secondary Hopf curve also returns the secondary rotation number to its value at the starting point, then relative extrema must exist. Thus, these scenarios will lead to Hopf bifurcations with angular degeneracies.

Differences in the figures depend on which side of the horn the second Takens–Bogdanov point appears, on which side of the secondary Hopf bifurcation curve the secondary invariant curves exist, and on which type of angular degeneracy is realized

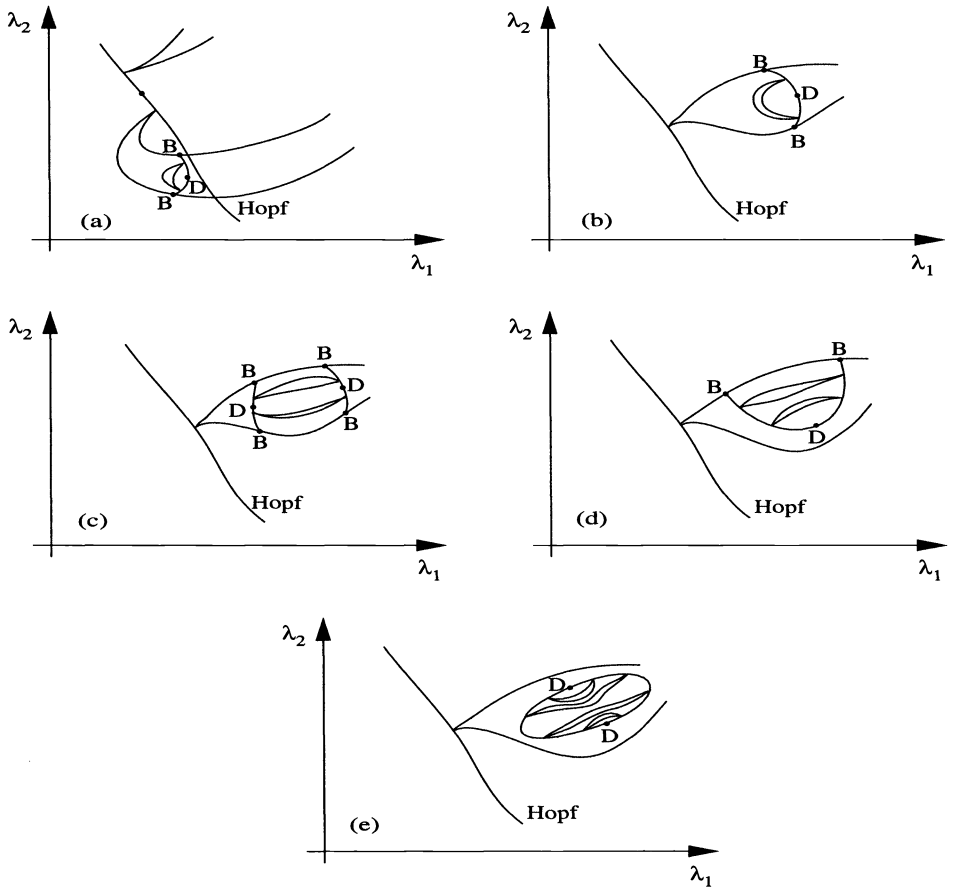


FIG. 5. *Angular degeneracies D due to Bogdanov points B.*

("banana" vs. "banana split"). Although there are no Takens–Bogdanov points in Fig. 5e, it could turn into Fig. 5d by "expanding" the secondary Hopf circle through a variation of an auxiliary parameter, for example, until the "top" angular degeneracy "hit" the saddle-node curves bounding the resonance horn. Other similar scenarios are also possible.

Figure 5a is an illustration of a pair of resonance horns which exist near a "Chenciner" point [Ch]. A Chenciner point is yet another degenerate Hopf bifurcation point: the Hopf bifurcations change between supercritical and subcritical at the Chenciner point. This is illustrated by the switch in the side of the primary Hopf bifurcation curve into which the primary resonance horns grow. As part of his thesis, Johnson [Jo] showed that there *necessarily* exist two Takens–Bogdanov points, one on each side of the primary resonance horn which "turns around," and a secondary Hopf curve which connects them. In this case, there *must* be a Hopf point with an angular degeneracy along that secondary Hopf bifurcation curve. The adaptive control application, which we describe next, has a bifurcation diagram with features similar to Fig. 5a.

3.2. The adaptive control application. Consider the problem of controlling the linear, discrete-time, single-input, single-output (SISO) plant with unknown, con-

stant coefficients (see the 1984 textbook by Goodwin and Sin [GS]):

$$y(t+1) = -\alpha_1 y(t) - \alpha_2 y(t-1) + \beta_0 u(t). \tag{5}$$

In designing the controller, a first-order reference model of (5) is assumed:

$$\hat{y}(t+1) = \hat{\alpha}_1(t) y(t) + \hat{\beta} u(t), \tag{6}$$

where $-\hat{\alpha}_1$ and $\hat{\beta}$ are estimates of the actual system parameters α and β. Thus, two sources of plant/reference-model error are introduced by the reference model: (1) the use of a first-order model (since $\hat{\alpha}_2 = 0, \alpha_2$ becomes a measure of the plant/reference-model *order* mismatch); (2) it is assumed that a good estimate of the gain of the manipulated variable (β_0) is known (thus, $\hat{\beta}$ is a constant). The objective of the controller $u(t)$ is to make the system follow the set point $y^*(t)$; inverting (6), the control law

$$u(t) = \frac{y^*(t+1) - \hat{\alpha}_1(t) y(t)}{\hat{\beta}}$$

is obtained. Choosing $y^*(t+1) = \text{constant} \neq 0$, it is possible to set $y^* = 1$ without loss of generality. The recursive identifier for α_1 is a scalar form of the projection algorithm of [GRC]:

$$\hat{\alpha}_1(t) = \hat{\alpha}_1(t-1) + y(t-1)\frac{y(t)-1}{c + y^2(t-1)}.$$

Defining $x(t+1) = y(t)$, the closed-loop system can be written as

$$\begin{aligned} x(t+1) &= y(t), \\ y(t+1) &= -\alpha_1 y(t) - \alpha_2 x(t) + \frac{\beta_0}{\hat{\beta}}\left(1 - \hat{\alpha}_1(t) y(t)\right), \\ \hat{\alpha}_1(t+1) &= \hat{\alpha}_1(t) + \frac{y(t)}{c + y^2(t)}(y(t+1) - 1), \end{aligned}$$

and after defining $a = -\alpha_1$, $b = -\alpha_2$, $k = \beta_0/\hat{\beta}$, and $z = a - k\hat{\alpha}_1$ the final form of the map G: $R^3 \mapsto R^3$

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} y \\ bx + k + zy \\ z - \frac{ky}{c+y^2}(bx + k + zy - 1) \end{pmatrix}$$

is derived. The system is characterized by three parameters. The small and positive constant c pertains to the estimation algorithm chosen; it is used to prevent division by zero in the estimator. In our calculations it was kept fixed at the representative value of $c = 0.1$. The second parameter, k, is a measure of the error in the assumption of the value of the gain of the manipulated variable ($k = 1$ implies no error), and finally, b is a measure of the plant/reference-model order mismatch ($b = 0$ implies no order error).

Bifurcation analysis reveals a Hopf bifurcation locus for the period-1 fixed point (it corresponds to the set point of the process) in the (k, b) parameter plane

$$b = b_h = -\frac{c+1}{c+2}.$$

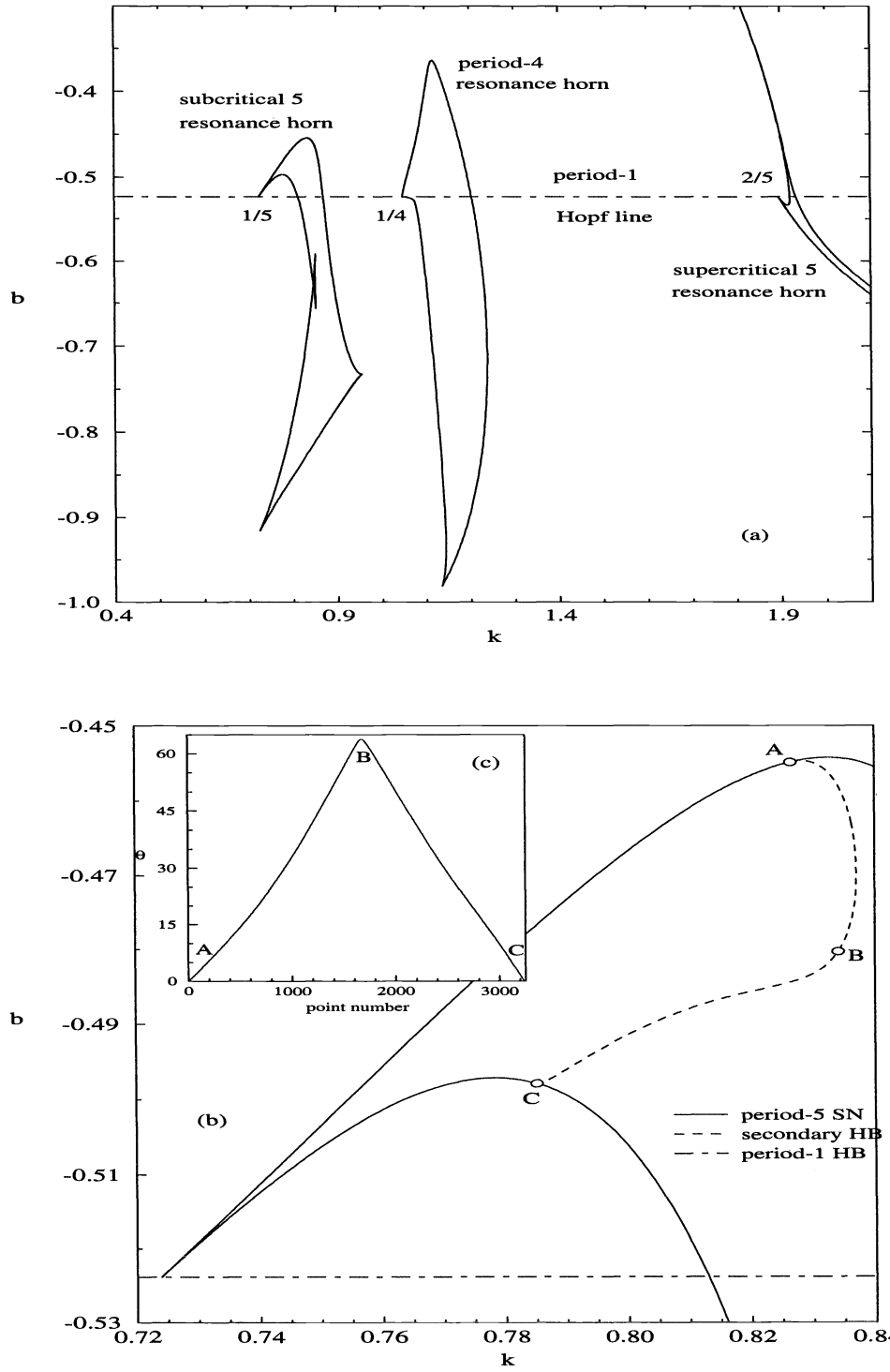


FIG. 6. *Adaptive control system*: (a) *parameter space*, (b) *local bifurcations inside the subcritical period-5 horn*, (c) *angle vs location on secondary Hopf curve.*

Along the Hopf-bifurcation locus, two complex eigenvalues are located on the unit circle (critical eigenvalues) while the third, real eigenvalue is given by

$$\lambda_1 = -b_h = \frac{c+1}{c+2}.$$

Since λ_1 lies well within the unit circle for our choice of $c = 0.1$, the dynamics are strongly contracting in the direction corresponding to λ_1. It is therefore expected that the system will behave in a fashion similar to a map of the plane in the neighborhood of the Hopf bifurcation. It can be easily shown that as k is varied along the Hopf-bifurcation line, the critical eigenvalues start at (-1,-1) at $k = 2.092857$ and then move monotonically over the entire unit circle, approaching (1,1) as $k \to 0$. As described, for example, in Corollary 2.4, primary resonance horns are expected to emanate from this line. This is confirmed in Fig. 6a; the details of the local bifurcations at the tip of the subcritical period-5 horn are shown in Fig. 6b [FAK], [Fr]. The continuation calculations were performed using AUTO86 by Doedel [Do], [DK] (and a real-time graphics interface for it by Dr. M. A. Taylor in our group).

On each side of this period-5 horn we observe a Takens–Bogdanov point (two eigenvalues of $D\mathbf{G}^5$ at one); they are marked A and C in Fig. 6b. As predicted by the theory, we were able to compute the secondary Hopf bifurcation curve inside the period-5 horn connecting the two Takens–Bogdanov points. Along this curve, the two relevant eigenvalues of the corresponding period-5 orbit "start" with zero argument (point A) and after reaching a maximum argument of about 63.8^o on the unit circle (point B—the angular degeneracy) they move back to zero argument at point C (Fig. 6c). Secondary resonance regions originate from this secondary Hopf curve. Figure 7 shows the banana-shaped secondary resonance horns associated with a 6th and 7th root of unity, when the eigenvalues are $\cos(\frac{2\pi}{6}) \pm i \sin(\frac{2\pi}{6})$ and $\cos(\frac{2\pi}{7}) \pm i \sin(\frac{2\pi}{7})$, respectively. The 6th root of unity is crossed twice along the AC curve (at points F and G) where $(k, b) \approx (0.8369, -0.4725)$ and $(0.8274, -0.48418)$, respectively. Similarly, the period-7 resonance horn opens and closes at points D and E on the AC curve, where $(k, b) \approx (0.8366, -0.4645)$ and $(0.8183, -0.48603)$, respectively. We have numerically traced the boundaries of the period-6 and period-7 resonance horns for G^5 in Fig. 7. (Period-6 (respectively, 7) for G^5 means period $5 * 6 = 30$ (respectively, $5 * 7 = 35$) for the original map G.) These secondary resonance horns both "open" and "close" on the secondary (i.e., period-5) Hopf bifurcation curve, suggesting that point B is a banana point rather than a banana split point. We would need to compute higher-order terms in the normal form on the center manifold, however, in order to be sure.

We note that in other examples with an angular degeneracy on a secondary Hopf curve the argument of the eigenvalue at the maximum point (that is, at the angular degeneracy) is only a couple of degrees instead of 63.8^o, as it is here. This is why this example was good for computing bananas: period 5×6 saddle-nodes are much easier to compute than, say, saddle-nodes of period 5×180.

4. Global bananas. All of our results to this point have been local in nature. Banana regions or banana split/parabolic regions have been shown to exist in arbitrarily small neighborhoods of a Hopf bifurcation point with an angular degeneracy. On the other hand, banana resonance regions seem to appear in our numerically computed bifurcation diagrams even relatively far from angular degeneracies. For example, the period-30 and 35 bananas of Fig. 7 seem relatively far from the angular degeneracy — far enough, at least, so that their shapes would not still be called

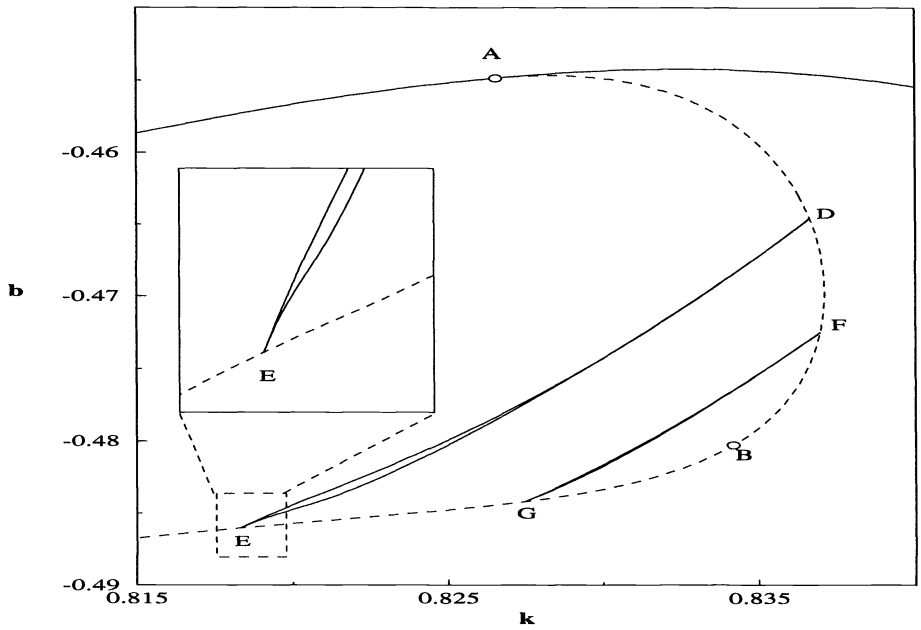


FIG. 7. *Period-30 and period-35 "bananas".*

parabolic. Also, in the schematic bifurcation diagrams of Fig. 5, all the secondary resonance regions are closed bananas, even in the case of Fig. 5c, where the two angular degeneracies are intended to be locally banana split points. It is even possible that the saddle-node curves, which bound the primary resonance regions of Fig. 5, if continued beyond the point where the diagrams stop, could "end" at a second cusp on another (or the same) primary Hopf bifurcation curve. We now give the following global banana result, where the existence of one p/q resonant Hopf point implies the existence of another.

THEOREM 4.1. *Let* $\mathbf{F}$ *be a* C^∞ *function from* $\mathbf{R}^2 \times \mathbf{R}^2 \to \mathbf{R}^2$ *which represents a two-parameter family of diffeomorphisms of the plane. Assume the following:*

1. *There is a Hopf bifurcation curve with a* p/q *resonant point,* $q \geq 3$, *which does not have an angular degeneracy.*

2. *The region of phase* $\times$ *parameter space where a* p/q *orbit exists is compact.*

Then there must exist another p/q *resonant Hopf point somewhere in that compact region of phase* $\times$ *parameter space. Both points are puncture points on the same component of least-period-q points (i.e., the component of the* p/q *resonance surface) in the phase* $\times$ *parameter space. (That is, the existence of one end of a banana implies a second end must also exist.)*

Proof. Theorem 2.3 tells us that the surface of period-q points near the assumed p/q resonant Hopf point is a punctured disk. The idea of the proof is to consider this surface globally in the phase $\times$ parameter space. It can be shown that the closure of the set of least-period-q points in phase $\times$ parameter space forms an orientable topological two-manifold. (In the simplest case, this manifold would be a topological sphere, but it might have some number of handles, as well.) All points on this manifold are least-period-q points under the map $(\mathbf{x}, \mu) \to (\mathbf{F}_\mu(\mathbf{x}), \mu)$, except possibly for isolated fixed points such as the p/q resonant Hopf point projecting to the (first) tip

of the resonance horn (if $q \geq 5$). The proof of the existence of the second fixed point on the resonance surface emanating from the first p/q resonant Hopf point is almost the same as the proof of Theorem 2 of [P2]. That theorem proves the existence of a p/q Hopf point on a p/q surface that emanates not from a first p/q Hopf point, but from "zero forcing amplitude" in a two-parameter family of maps of the plane generated by return maps of a periodically forced planar oscillator. The p/q surface for a forced oscillator "naturally" has an invariant circle as a boundary component; the map restricted to this invariant circle is a rigid rotation by p/q. To convert our situation to that of [P2], we need to replace the first p/q Hopf point with a boundary circle on which the map is a rotation by p/q. But this is easily done by "blowing up" the p/q resonant Hopf point (extending the phase space in polar coordinates to $r = 0$). The proof of the existence of the second p/q Hopf point then follows from [P2]. □

We next present a corollary which describes conditions under which a whole collection of secondary global banana regions will exist.

COROLLARY 4.2. *Let $\mathbf{F}_\mu$ be a generic two-parameter family of diffeomorphisms of $\mathbf{R}^2$. Assume the following:*

1. *There is a p/q resonance surface in the phase $\times$ parameter space resulting from a (primary) p/q resonant Hopf bifurcation.*
2. *The p/q resonance surface includes two Takens–Bogdanov points for the qth iterate of the map; the two Takens–Bogdanov points are connected by a secondary Hopf bifurcation curve (also along the p/q resonance surface).*
3. *Along the secondary Hopf curve the argument of the neutral eigenvalue of $D\mathbf{F}_\mu^q$ has a single local extremum, say $2\pi\omega_0$.*
4. *There is no other secondary Hopf curve on the p/q resonance surface.*
5. *All secondary periodic point surfaces emanating from the secondary Hopf curve are contained in a compact region of phase $\times$ parameter space.*

Then, for every $m/n \in (0, \omega_0)$, the period-qn surface emanating out of the m/n secondary Hopf point must connect to the period-qn resonance surface emanating from the unique m/n Hopf point on the secondary Hopf curve on the other side of the local extremum. (Thus, all resonance regions emanating from the secondary Hopf curve are globally closed bananas.)

Proof. The hypotheses of Theorem 4.1 are satisfied for each $m/n \in (0, \omega_0)$, so a second m/n Hopf point must exist. The assumptions of a single local extremum and no other secondary Hopf curves imply that there is only one "appropriate" point. This point, therefore, is where the other end of the global banana must be. □

Note: It seems that all primary resonance horns near and on one side of a Chenciner point on a Hopf bifurcation curve satisfy the hypotheses of Corollary 4.2. This would give us an infinite collection of primary resonance horns, each having its own infinite collection of global bananas.

5. Conclusions and comments. Although the parametric degeneracy we studied in this paper was specifically along a Hopf bifurcation curve, *any* parametric degeneracy (with respect to parameters in a universal unfolding of a local bifurcation) can be thought of, in its simplest form, as merely a local "folding in half" of the degenerate parameter space, in order to map it to the universal (nondegenerate) parameter space. We could, for example, have included the strongly resonant cases in Theorems 2.3 and 2.5 and their corollaries, even though the projections of the resonance surfaces near the strongly resonant Hopf points to the (nondegenerate) parameter space are not necessarily cusps.

It might be useful to write explicit conditions in terms of the original map to determine (a) an angular degeneracy and (b) the type: banana vs. banana-split (harder, since higher-order terms are required). We found it much easier to verify conditions of the theorem by numerically computing arguments of eigenvalues along the Hopf curve, as we did for the adaptive control application to produce Fig. 6c, than by computing a normal form (especially when needing to use a center manifold).

We point out that the global results (Theorem 4.1 and Corollary 4.2) are very much dependent on the phase space being two-dimensional. The fixed-point theorem from [P2] quoted in the proof of Theorem C applies only in that setting. On the other hand, we expect local results in higher dimensions to be preserved by use of a center manifold. Note that Corollary 4.2 does not exclude the possibility of "nonbanana" resonance regions which do not emanate from the secondary Hopf curve. For example, if the local banana-split horn "partners" connect to form a global banana, we would expect the local parabolic regions to also connect, forming global annuli, projections of tori from the phase $\times$ parameter space. This scenario can be imagined by extending the two p_1/q_1 horns in Fig. 1b until they connect, forming a global banana; the global p_0/q_0 region would then likely be an annulus.

We caution our readers that knowing the complete structure of resonance regions for a family of maps does not necessarily mean we have a complete bifurcation classification, even locally in a neighborhood of a nondegenerate Hopf bifurcation point. We do know that all maps on the side of the Hopf bifurcation curve without the invariant circle, including all those on the curve itself, are locally topologically equivalent. We also know that on the side with the invariant curve, the parameter space must be divided at least into the following equivalence classes: the interiors of each resonance region (circles in resonance), each boundary of each resonance region (circles in resonance with saddle-node orbits), and curves "parallel" to the resonance regions along which the corresponding maps restricted to the invariant circle are conjugate to a rigid rotation with an irrational rotation number. What is missing is a guarantee that all the maps in a given resonance region are equivalent. Corollary 2.4 comes close to giving this guarantee: the existence of a *single* attracting/repelling pair of periodic orbits as stated in part 3 of Corollary 2.4 implies that all maps corresponding to parameter values in the interior of a resonance horn and close enough to the tip *are* topologically equivalent. This may not, however, imply that this uniqueness of equivalence classes within a single resonance region can be extended to hold for *all* resonance regions in a fixed neighborhood (not depending on p/q) of a Hopf bifurcation point. This is why in Corollary 2.6, where we make a claim about the shapes of resonance regions "for all p/q sufficiently close to ω_0," we were unable to claim, as we did in Theorem 2.3, Corollary 2.4, and Theorem 2.5, that there exists a single pair of period-q orbits inside the corresponding p/q resonance region.

Even if a complete local classification could be established, no such claim could ever be made about the global bananas being the complete bifurcation diagram. Check [ACHM], for example, to see a variety of possible further subdivisions of a single resonance region into further equivalence classes. These further subdivisions are possible in part because, away from the Hopf curve, as well as near strong resonances, the invariant circle which is born in the Hopf bifurcation may break. This allows nonuniqueness of rotation numbers, which in turn allows resonance regions to overlap. Near strong resonances, in fact, they *must* overlap, because of global manifold crossings which imply the existence of an infinite number of periodic orbits for fixed parameter values.

There are an additional number of related questions we have not addressed in this paper: (a) No upper bound is given on the number of resonant Hopf points which may exist on a given two-manifold of period-q points; more than two could certainly exist. We conjecture that Lefschetz index theory could be used to show that the fixed points should generically come in pairs, having indices plus and minus one, respectively. "Mutant" bananas, with 4 tips, for example, could easily be constructed by parameter space surgery on a family having a banana with 2 tips! (b) Cusp points (saddle-nodes with a higher-order degeneracy—not to be confused with cusps at resonant Hopf points) may also appear along the saddle-node boundaries of resonance regions. They usually appear in pairs, as well, such as on the left-hand side of the subcritical period-5 resonance horn in Fig. 6a. Work in progress further describes these pairs of cusps [MP]. (c) Finally, finding examples that would exhibit all schematic scenarios pictured in Figs. 5a–5e remains, to our knowledge, an open problem.

6. Appendix: Proofs.

6.1. Proof of Theorem 2.3. As indicated in §2.2, many of the arguments in this section are adaptations of arguments which Arnold [Ar] uses for $q = 4$. We also note that the symmetry of equation (2) implies that $\mathbf{f}^q(\mathbf{z}) = \mathbf{z}$ is equivalent to $\mathbf{f}(\mathbf{z}) = e^{2\pi ip/q}\mathbf{z}$. The latter equation is easier to use for verifying property 1 of Theorem 2.3, but more difficult to generalize to Corollary 2.4, where the symmetry is not present. So we stick to solving $\mathbf{f}^q(\mathbf{z}) = \mathbf{z}$.

Property 1.

We start with the following lemmas.

LEMMA 6.1. *Assume $q \geq 1$ and*

$$\mathbf{f}(\mathbf{z}) = \mu(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + \cdots + B\overline{\mathbf{z}}^{q-1} + \cdots), \tag{7}$$

where the omitted terms are all $O(|\mathbf{z}|^{q+1})$, except those of the form $\mathbf{z}^j\overline{\mathbf{z}}^{j-1}, 3 \leq j \leq \frac{q+1}{2}$, which are $O(|\mathbf{z}|^5)$. Then

$$\mathbf{f}^n(\mathbf{z}) = \mu^n\left(\mathbf{z} + A\left(\sum_{k=0}^{n-1}|\mu|^{2k}\right)\mathbf{z}^2\overline{\mathbf{z}} + \cdots + B\left(\sum_{k=0}^{n-1}|\mu|^{-2k}\overline{\mu}^{kq}\right)\overline{\mathbf{z}}^{q-1}\right) + \cdots, \tag{8}$$

where the omitted terms are all $O(|\mathbf{z}|^{q+1})$, except those of the form $\mathbf{z}^j\overline{\mathbf{z}}^{j-1}, 3 \leq j \leq \frac{q+1}{2}$, which are $O(|\mathbf{z}|^5)$.

Proof. Direct calculation and induction on n. □

LEMMA 6.2. *Assume the $C^k, k \geq 2$ family of C^∞ maps, $\mathbf{f}_{(\rho,\alpha)}$, is defined by*

$$\mathbf{f}_{(\rho,\alpha)}(\mathbf{z}) = e^{2\pi ip/q}e^{\rho+i\alpha}(\mathbf{z} + A(\rho,\alpha)\mathbf{z}^2\overline{\mathbf{z}} + \cdots + B(\rho,\alpha)\overline{\mathbf{z}}^{q-1} + \cdots), \tag{9}$$

where the omitted terms are as in Lemma 6.1. Let $A = A(0,0), B = B(0,0), \mathbf{z} = re^{i\theta}$. Assume $A \neq 0, B \neq 0$, and $q \geq 2$. Then the least-period-q points near $(\rho, \alpha, \mathbf{z}) = (0,0,\mathbf{0})$ are given by the solutions of the equation

$$\rho + i\alpha + Ar^2 + \cdots + Br^{q-2}e^{-qi\theta} + \cdots = 0, r \neq 0, \tag{10}$$

where the θ-independent omitted terms are $O(r^4, \rho^2, \alpha^2, \alpha\rho, \rho r^2, \alpha r^2)$, and all other omitted terms are $O(r^q, \rho r^{q-2}, \alpha r^{q-2})$.

Proof. Period-q points satisfy $\mathbf{f}^q(\mathbf{z}) - \mathbf{z} = \mathbf{0}$. Use Lemma 6.1 to compute $\mathbf{f}^q(\mathbf{z})$, and expand $A(\rho,\alpha), B(\rho,\alpha), \mu^q, \mu^q - 1, \sum_{k=0}^{q-1}|\mu|^{2k}$, and $\sum_{k=0}^{q-1}|\mu|^{-2k}\overline{\mu}^{kq}$, with $\mu =$

$e^{2\pi ip/q}e^{\rho+i\alpha}$ to get power series in ρ and α. Then substitute $\mathbf{z} = re^{i\theta}$ and divide through by $qre^{i\theta}$. (Dividing by r eliminates only the origin, which is a fixed point.) It can be shown (Proposition 3.2 in [CMY] or Theorem 1, part A in [P2]) that solutions to equation (10) can only have a least period of q or 1. Since period-1 points are only at $\mathbf{z} = 0$, then period-q points with $r \neq 0$ must be *least* period-q points. □

Although we cannot solve (10) for the period-q points as a function of the parameters ρ and α, we can solve for the parameters as a function of the phase variables r and θ. By the implicit function theorem on (10), this is apparently

$$\rho + i\alpha = -Ar^2 - \cdots - Br^{q-2}e^{-qi\theta} - \cdots. \tag{11}$$

By choosing r small enough, say less than r_0, we can be sure that Ar^2 dominates all omitted θ-independent terms, and $Br^{q-2}e^{-qi\theta}$ dominates all θ-dependent terms.

For $0 < r \leq r_0, 0 \leq \theta < 2\pi$ equation (11) is an explicit parametrization of the punctured disk which is the least-period-q surface. Adding r=0, corresponding to the resonant Hopf point, fills in the puncture point in the disk. This completes the proof of Property 1.

Property 2. We now use equation (11) to determine the region in parameter space to which this disk projects. Ignoring the omitted higher-order terms, which are the same as for equation (10) in the statement of Lemma 6.2, and, for now, the r^{q-2} term, we get $\rho + i\alpha = -Ar^2 + \cdots$, or equating real and imaginary parts, respectively,

$$\rho = -A_1 r^2; \alpha = -A_2 r^2. \tag{12}$$

Eliminating r gives

$$\alpha = \frac{A_2}{A_1}\rho, \frac{\rho}{-A_1} > 0. \tag{13}$$

That is, to the lowest-order terms in ρ and α, the parameter values for which period-q points exist trace out a ray in the parameter space from the origin in the direction of $(-A_1, -A_2)$ as r increases from 0.

If we now include the θ-dependent term from equation (11), $Br^{q-2}e^{-qi\theta}$, we see that for a fixed value of r, and letting θ vary from 0 to 2π, a circle in the parameter space is swept out (q times), having center at $(\rho, \alpha) = (-A_1 r^2, -A_2 r^2)$ and radius $|B|r^{q-2}$. When $q \geq 5$, sweeping out all such circles for small r covers a horn-shaped region in the parameter space. See Fig. 8 where we have drawn two such circles and the corresponding horn for a resonance region. Since the distance from the origin of these circles varies with r^2 and the width of the horn varies with r^{q-2}, the two sides of the horn are tangent of order $\frac{q-2}{2}$.

This completes the proof of property 2 of Theorem 2.3, but as a heuristic comment, we note that the terms on the right-hand side of equation (11), including those not explicitly written, can be separated into θ–independent and θ–dependent terms. If we considered all θ–independent terms, the analogue of equation (12) would be a semi-infinite curve instead a straight ray. This curve we call the "center of the resonance horn." (This would be well defined if the equation were completely in normal form—all nonresonant terms eliminated. This, however, would bring up the question of convergence of an infinite sequence of coordinate changes. Our wish to avoid this technicality is why these comments are merely heuristic.) The center of the resonance horn is still, of course, tangent to the vector $(-A_1, -A_2)$ at the origin. Now adding the $Br^{q-2}e^{-qi\theta}$ term will cause parameter space circles to be swept out as θ varies with r

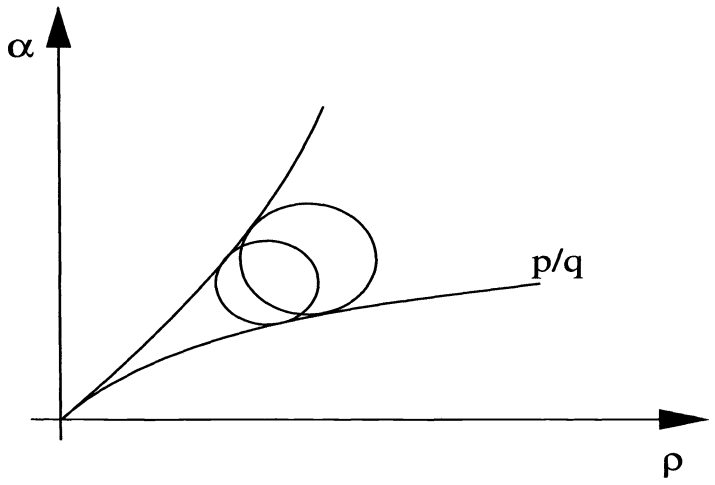


FIG. 8. *"Sweeping out" the resonance horn.*

held fixed. The centers of the circles are on the center of the resonance horn. Finally, including the higher-order θ–dependent terms will cause the circles which are swept out as θ varies to be slightly deformed. The horn sides are still tangent to $(-A_1, -A_2)$ at the origin, and the order of tangency is still $\frac{q-2}{2}$. Thus the only parameter values near $(\rho, \alpha) = (0, 0)$ for which period-q points near $\mathbf{z} = \mathbf{0}$ can exist are inside and on the boundary of the described resonance horn in the parameter space.

Property 3. (As in Arnold [Ar] for q=4.) From the terms that *do* explicitly appear in equation (11), it is apparent that any point in the interior of the horn lies on exactly two distinct circles, each circle corresponding to a different value of r and having its respective center at $(-A_1 r^2, -A_2 r^2)$. See Fig. 8 again. As θ varies from 0 to 2π, each circle is traced out q times (in the negative angular direction). When included, the higher-order terms do not qualitatively affect this result. Thus there are q different phase points which correspond to the same parameter value on the circle. In fact, together these q phase points form one complete period-q orbit. Thus each parameter value on the interior of the horn has two distinct period-q orbits for the associated map. That one is a saddle and the other a node is verifiable using techniques similar to those used by Arnold for $q = 4$ (§35J in [Ar]).

6.2. Proof of Corollary 2.4. The normal form theorem assures us that the original equation can be brought into the form of equation (1). We would like to make a change of parameters from $\boldsymbol{\mu}$ to (ρ, α) where the relationship between them has already been defined by $\frac{2\pi p}{q} + \alpha = \phi$ and equation (1). This is possible if the change of parameters, which we will call $\mathbf{h}$, is nonsingular at $\boldsymbol{\mu}_0$. Equivalently, we must have the vectors $\nabla_\mu \rho(\boldsymbol{\mu}_0)$ and $\nabla_\mu \alpha(\boldsymbol{\mu}_0)$ being independent, which we do because we assumed the absence of a parametric degeneracy (defined in the introduction).

Renaming $\mathbf{f}_\mathbf{u} = \mathbf{f}_{\mathbf{h}^{-1}(\rho,\alpha)}$ to be $\mathbf{f}_{\rho,\alpha}$ brings the equation into the form

$$(14)\quad \mathbf{f}_{(\rho,\alpha)}(\mathbf{z}) = e^{2\pi i p/q} e^{\rho + i\alpha}(\mathbf{z} + A(\mathbf{h}^{-1}(\rho,\alpha))\mathbf{z}^2\bar{\mathbf{z}} + \cdots + B(\mathbf{h}^{-1}(\rho,\alpha))\bar{\mathbf{z}}^{q-1} + \cdots).$$

Now we push through the conclusions and proofs of Theorem 2.3 using the family of equation (14), which is a generalization of our model family of equation (2). The lemmas used to prove Theorem 2.3 were actually proved already in the more general

form of equation (14). Compare equation (14) with equation (9) in Lemma 6.2, in particular. Since the results of Theorem 2.3 hold for equation (14), both results of the corollary now follow directly from the fact that the function $\mathbf{h}^{-1}$ is a nonsingular C^∞ map from the (ρ, α) parameter plane to the μ parameter plane.

6.3. Proof of item 1 of Theorem 2.5. Rewrite equation (3), replacing ω_0 with $p/q + \alpha_0/2\pi$:

$$\mathbf{f}_{(\rho,\tau)}(\mathbf{z}) = e^{2\pi i p/q} e^{\rho + i(\alpha_0 + c_1\rho + c_2\tau^2)}(\mathbf{z} + A\mathbf{z}^2\overline{\mathbf{z}} + B\overline{\mathbf{z}}^{q-1}). \tag{15}$$

This is the same as the equation for the nondegenerate Hopf bifurcation (equation (2)), but with $\alpha_0 + c_1\rho + c_2\tau^2$ replacing α. Since the nondegenerate analysis was valid for α sufficiently small, the same analysis will hold for $\alpha_0 + c_1\rho + c_2\tau^2$ sufficiently small. We treat α_0 as a third parameter which is small if ω_0 is sufficiently close to p/q. Therefore we first consider the five-dimensional phase $\times$ parameter space, and then obtain the theorem by restricting to an "α_0 = small constant" slice.

The least-period-q set, analogous to equation (10), becomes

$$\rho + i(\alpha_0 + c_1\rho + c_2\tau^2) + Ar^2 + \cdots + Br^{q-2}e^{-qi\theta} + \cdots = 0, r \neq 0. \tag{16}$$

Thinking of this complex equation as two scalar equations, we see that the Jacobian with respect to ρ and r^2 at $(r^2, \theta, \rho, \tau, \alpha_0) = (0, \theta, 0, 0, 0)$ is $A_2 - c_1A_1$, which was assumed to be nonzero. So we can solve locally for ρ and r^2 as a function of θ, τ, and α_0. This would seem to indicate the period-q surface is always locally a cylinder: $\theta \in \mathbf{S}^1, \tau \in$ an interval. But the circle swept out as θ varies for a fixed value of r^2 collapses to a point (a fixed point of the map) when $r^2 = 0$ and doesn't exist if $r^2 < 0$. So we must determine the topology of the least-period-q set by determining the values of τ which correspond to $r^2 > 0$ and to $r^2 = 0$. This is what we proceed to do.

By ignoring the θ-dependent terms for now, and eliminating r from equation (16), we obtain an expression analogous to equation (13) for the "center" of our p/q resonance horn to lowest order in the three small parameters ρ, τ, and α_0: $\alpha_0 + c_1\rho + c_2\tau^2 = \frac{A_2}{A_1}\rho, \frac{\rho}{-A_1} > 0$. This is equivalent to

$$\tau^2 = \frac{A_2 - c_1A_1}{c_2A_1}\left(\rho - \alpha_0\frac{A_1}{A_2 - c_1A_1}\right), \frac{\rho}{-A_1} > 0. \tag{17}$$

Treating α_0 as small, nonzero, and fixed, we see that there are actually eight cases, all pieces of parabolas, depending on the signs of $(A_2 - c_1A_1)/(c_2A_1), (\alpha_0A_1)/(A_2 - c_1A_1)$, and A_1. We have sketched the four cases assuming $A_1 < 0$ in Fig. 9. The dashed lines are included in the diagram merely for reference—they are the part of the parabola excluded by $\frac{\rho}{-A_1} > 0$, which corresponds to the side of the Hopf curve without the invariant circle. If A_1 were positive, we would get four similar cases, each a reflection across the τ axis of one of the Fig. 9 cases. (It might be useful to compare Fig. 9a with Figs. 3a, $4d_1$, and the horns near D_1 in Fig. 1b; Fig. 9b with Figs. 3b, $4d_2$, and the horns near D_2 in Fig. 1b; and Fig. 9c with Figs. 3c, $4d_2$, and the horns near D_2 in Fig. 1b.)

We choose an appropriate neighborhood of phase $\times$ (ρ, τ) space by first restricting (ρ, τ, α_0) to a small enough neighborhood of the origin so that the terms explicitly written in equation (16) dominate the (higher-order) unwritten terms. We can choose this neighborhood as a cube with sides at $\rho = \pm\overline{\rho}, \tau = \pm\overline{\tau}, \alpha_0 = \pm\overline{\alpha_0}$ by making small enough choices for $\overline{\rho}, \overline{\tau}$, and $\overline{\alpha_0}$. In cases (a) and (b), we further restrict $\overline{\alpha_0}$,

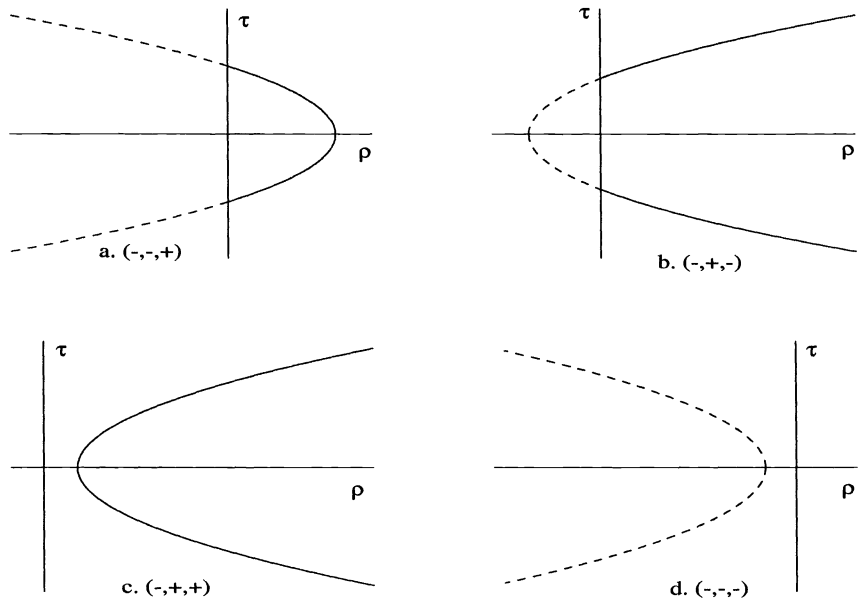


FIG. 9. *"Centers" of resonance regions near a Hopf bifurcation with an angular degeneracy. Labels are for the signs of* $A_1, (A_2 - c_1A_1)/(c_2A_1)$, *and* $(\alpha_0 A_1)/(A_2 - c_1A_1)$, *respectively.*

if necessary, so that $\sqrt{|\overline{\alpha_0}/c_2|} < \overline{\tau}/2$. This ensures that in an "α_0 = constant" slice of our three-dimensional space, the banana "tips" on the τ axis are included in the neighborhood.

From Fig. 9, it is now apparent that the center curve(s) can be parametrized in the four respective cases (a), (b), (c), and (d) by $\tau \in$

(a) (τ_-, τ_+),
(b) $(-\overline{\tau}, \tau_-) \cup (\tau_+, \overline{\tau})$,
(c) $(-\overline{\tau}, \overline{\tau})$,
(d) the empty set,

where from equation (17), $\tau_\pm = \pm\sqrt{|\alpha_0/c_2|}$.

Reintroducing the θ-dependent terms from equation (16) and varying $\theta \in [0, 2\pi)$ gives a parametrization of the least-period-q surface as the product of the appropriate set of the above four for τ with the unit circle. The puncture points corresponding to $r^2 = 0$ in cases (a) and (b) are at $\tau_\pm$. This gives us the twice-punctured sphere for case (a), the two punctured disks for case (b), the cylinder for case (c), and the empty set for case (d). Including the neglected higher-order terms does not change the topology of these sets.

For the projections to the (ρ, τ) parameter space, we fix τ (α_0 is already fixed) and let θ vary from zero to 2π. This traces out a closed curve restricted to τ = constant in the parameter space. Unless the θ-dependent terms all vanish (including all higher-order terms), the closed curve covers a positive length line segment which provides the "thickening" of the respective center lines into regions and establishes the shapes of the respective resonance regions. If the coefficient $B \neq 0$, the length of the line segment varies with r^{q-2} while the distance from the τ axis varies with r^2, establishing the order of tangency at the tips.

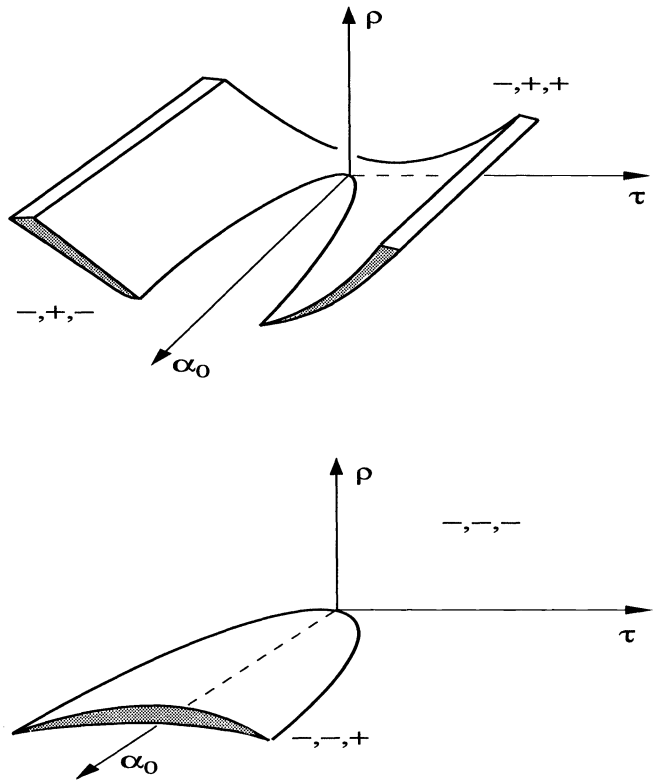


FIG. 10. *Resonance regions in a three parameter space near a Hopf bifurcation with an angular degeneracy. Labels are for the signs of* $A_1, (A_2 - c_1A_1)/(c_2A_1)$, *and* $(\alpha_0 A_1)/(A_2 - c_1A_1)$, *respectively.*

For further illustration we have sketched the resonance regions in the three-parameter space (ρ, τ, α_0) in Fig. 10 for two distinct cases. The sign of A_1 is assumed to be negative in both cases; the sign of $(A_2 - c_1A_1)/(c_2A_1)$ is assumed to be positive in the first case, negative in the second. The sign of the third quantity, $(\alpha_0 A_1)/(A_2 - c_1A_1)$, is determined by the sign of α_0, which is one of the parameters in the figure.

6.4. Proof of Corollary 2.6. Restrict a neighborhood of $(\mathbf{x}_0, \boldsymbol{\mu}_0)$ in phase $\times$ parameter space so that it contains no strongly resonant Hopf points. Choose a p/q with the condition that there is a Hopf point in the restricted neighborhood with eigenvalues $e^{\pm 2\pi i p/q}$. Delete from this neighborhood any of the Hopf points with eigenvalues $e^{\pm 2\pi i r/s}$ with $s < q$. On this deleted, restricted neighborhood, we can change variables to write the equations in the form of equation (1). This defines the functions $\rho(\boldsymbol{\mu}_0)$ and $\phi(\boldsymbol{\mu}_0)$.

We define a change of parameters from $\boldsymbol{\mu}$ to (ρ, τ) where $\rho(\boldsymbol{\mu})$ is defined by equation (1) and $\tau(\boldsymbol{\mu})$ is a linear variable with respect to the $\boldsymbol{\mu}$ parameter space in a direction perpendicular to $\nabla_{\mu}\rho(\boldsymbol{\mu}_0)$. This makes $\nabla_{\mu}\tau(\boldsymbol{\mu}_0)$ and $\nabla_{\mu}\rho(\boldsymbol{\mu}_0)$ independent vectors and ensures that the parameter change is nonsingular, and therefore a local C^{∞} diffeomorphism.

The normal form of equation (1) can now be rewritten, after replacing $\mathbf{h}^{-1}(\rho, \tau)$

with (ρ, τ), as

$$\mathbf{f}_{(\rho,\tau)}(\mathbf{z}) = e^{\rho + i\phi(\rho,\tau)}(\mathbf{z} + A(\rho,\tau)\mathbf{z}^2\bar{\mathbf{z}} + \cdots + B(\rho,\tau)\bar{\mathbf{z}}^{q-1} + \cdots), \tag{18}$$

where the series expansion of ϕ is $\phi(\rho,\tau) = \omega_0 + c_1\rho + c_2\tau^2 +$ higher-order terms in ρ and τ. The constant term ω_0 is determined by the eigenvalues at the bifurcation point $\boldsymbol{\mu}_0$; the τ term is absent because of the angular degeneracy; c_1 is nonzero because we assumed $\nabla_\mu(\alpha(\boldsymbol{\mu}_0)) \neq 0$; c_2 is nonzero because we assumed that $\frac{d^2}{ds^2}\alpha(\boldsymbol{\mu}(s))|_{s=0} \neq 0$. Except for the higher-order terms in $\phi(\rho,\tau)$, this family is the same as our model degenerate family of equation (3). The proof of Theorem 2.5 still works for the family in equation (18) because the higher-order terms in the expansion of ϕ contribute only to terms already considered as higher order in equation (16).

The fact that the change of parameters between (ρ,τ) and $\boldsymbol{\mu}$ is a local diffeomorphism and the fact that no period-q points can exist arbitrarily close to any of Hopf points deleted from our neighborhood completes the proof.

Acknowledgments. We are grateful to R. P. McGehee for helpful discussions and suggestions.

REFERENCES

[Ar] V. I. ARNOLD, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.

[ACHM] D. G. ARONSON, M. A. CHORY, D. G. HALL, R. P. MCGEHEE, *Bifurcations from an invariant circle for two-parameter families of maps of the plane: a computer assisted study*, Comm. Math. Phys., 83 (1983), pp. 303–354.

[Bo] R. I. BOGDANOV, *Versal deformation of a singularity of a vector field in the plane in the case of zero eigenvalues*, Trudy Seminara Imeni I. G. Petrovskogo, 2 (1976), pp. 23–36. (In Russian.) English translation: Selecta Math. Sovietica, 1 (1981), pp. 389–421.

[Ch] A. CHENCINER, *Bifurcation de points fixes elliptiques. II. Orbites periodiques et ensembles de Cantor invariants*, Invent. Math., 80 (1985), pp. 81–106.

[CMY] S. N. CHOW, J. MALLET-PARET, AND J. A. YORKE, *A periodic orbit index which is a bifurcation invariant*, in Geometric Dynamics (Lecture Notes in Mathematics 1007), J. Palis, ed., Springer-Verlag, 1983, pp. 109-131.

[Do] E. J. DOEDEL, *AUTO: a program for the automatic bifurcation analysis of autonomous systems*, Congr. Numer., 30 (1981), pp. 265–284.

[DK] E. J. DOEDEL AND J. P. KERNEVEZ, *AUTO: Software for continuation and bifurcation problems in ordinary differential equations (including the AUTO User Manual)*, Tech. Report, Applied mathematics, California Institute of Technology, 1986.

[Fr] C. E. FROUZAKIS, *Dynamics of Systems under Control: Quantifying Stability*, Ph.D. Thesis, Princeton University, 1992.

[FAK] C. E. FROUZAKIS, R. A. ADOMAITIS, AND I. G. KEVREKIDIS, *Resonance phenomena in an adaptively-controlled system*, International Journal of Bifurcation and Chaos, 1 (1991), pp. 83–106.

[GRC] G. C. GOODWIN, P. J. RAMADGE, AND P. E. CAINES, *Discrete-time multivariable adaptive control*, IEEE Trans. Automat. Control., 25 (1980), pp. 449–456.

[GS] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction, and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[GH] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, in Applied Mathematical Sciences, Vol. 42, Springer-Verlag, New York, 1983.

[Jo] J. R. JOHNSON, *Some Properties of a Three-Parameter Family of Diffeomorphisms of the Plane, Near a Transcritical Hopf Bifurcation*, Ph.D. Thesis, University of Minnesota, 1985.

[Mc] R. P. MCGEHEE, personal communications, 1987-88.

[MP] R. P. MCGEHEE AND B. B. PECKHAM, *Arnold flames and resonance surface folds*, Int. J. of Bifurcation and Chaos, submitted.

[P1] B. B. PECKHAM, *The Closing of Resonance Horns for Periodically Forced Oscillators*, Ph.D. Thesis, University of Minnesota, 1988.

[P2] B. B. PECKHAM, *The necessity of the Hopf bifurcation for periodically forced oscillators with closed resonance regions*, Nonlinearity, 3 (1990), pp. 261–280.

[Ru] D. RUELLE, *Elements of Differentiable Dynamics and Bifurcation Theory*, Academic Press, Inc., New York, 1989.

[Ta] F. TAKENS, *Forced oscillations and bifurcations*, Applications of Global Analysis, Vol. 3, Communications of the Mathematical Institute Rijksuniversiteit, Utrecht, 1974, pp. 1–59.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 218–237, January 1995

011

TRACE FORMULAS AND THE BEHAVIOUR OF LARGE EIGENVALUES*

VASSILIS G. PAPANICOLAOU†

Abstract. Let $\mu_1, \mu_2, \ldots, \mu_n, \ldots$ be the Dirichlet spectrum of the operator $-d^2/dx^2 + q$ acting on $L^2(0,b)$. In the special case where $q \equiv 0$, $\mu_n = \pi^2 n^2/b^2$. In the early 1950s Gelfand, Levitan [*Dokl. Akad. Nauk SSSR*, 88 (1953), pp. 593–599], and others discovered the asymptotic formula

$$\mu_n = \frac{\pi^2 n^2}{b^2} + \frac{1}{b}\int_0^b q(x)dx + O(n^{-2})$$

and the trace formula

$$\sum_n \left[\mu_n - \frac{\pi^2 n^2}{b^2}\right] = \frac{q(0)+q(b)}{4}, \qquad \text{provided that } \int_0^b q(x)dx = 0,$$

where $q \in C^2[0,b]$. These are beautiful formulas with many applications, for example in solving inverse problems.

Inspired by the above formulas, this paper obtains some results involving the spectra of two self-adjoint operators L and L_0 (where L can be thought as a perturbation of L_0). The following cases are considered:

(i) $L_0 = -d^2/dx^2$ and $L = L_0 + Q(x)$, with $Q(x)$ being an $r \times r$ real symmetric matrix (thus L_0 and L act on vectors $u = (u_1, \ldots, u_r)$, $u_j \in L^2(0,b)$);

(ii) $L_0 = (-\Delta)^m$ with Dirichlet boundary conditions and $L = L_0 + q$ acting on $L^2(0,b)$ or $L^2(D)$, where $D = (0,b_1) \times (0,b_2)$. The fact that ∂D has corners, thus it is not smooth, plays an essential role.

Some remarks are also made for the case where $L_0 = -\Delta$ with Neumann boundary conditions and $L = -\Delta$ with boundary conditions of the third kind (Robin).

Key words. higher-order Schrödinger operator, heat kernel, trace formula, diophantine number

AMS subject classifications. 35J10, 35J40, 35K35, 47A70

Introduction.

Consider the classical eigenvalue problem

$$Lu(x) = -u''(x) + q(x)u(x) = \mu u(x), \qquad u(0) = u(b) = 0,$$

where q is a smooth function on $[0,b]$. Problems of this type have been studied for the last 150 years. They have very rich mathematical theory and numerous applications. Two of the most famous results concerning the spectrum $\{\mu_n\}_{n=1}^{\infty}$ of the above problem are the *asymptotic formula*

$$\mu_n = \frac{\pi^2 n^2}{b^2} + \frac{1}{b}\int_0^b q(x)dx + O(n^{-2})$$

and the *trace formula*

$$\sum_n \left[\mu_n - \frac{\pi^2 n^2}{b^2}\right] = \frac{q(0)+q(b)}{4} \quad \text{provided that } \int_0^b q(x)dx = 0.$$

*Received by the editors January 14, 1992; accepted for publication (in revised form) June 8, 1993. This research was supported by National Science Foundation grant DMS-9011641.

†Department of Mathematics and Statistics, Wichita State University, Wichita, Kansas 67260-0033.

These formulas can be thought of as comparisons of the spectrum of L with the set $\{\nu_n = \pi^2 n^2/b^2\}_{n=1}^{\infty}$, namely the spectrum of the "unperturbed" operator $L_0 u = u''$. Their most important application is in solving inverse problems, namely, given some spectral-related data, how do we reconstruct the (now unknown) function q. The asymptotic formula, for example, gives a necessary condition for the existence of a q, while the trace formula, in certain formulations of the inverse problem, actually takes part in the explicit construction of q. In the periodic case, for instance (see [C]), and also in the case where two spectra are known (see [B] for a variant of this case), the solution of the inverse problem is achieved by deriving differential (evolution) equations of the eigenvalues with respect to the interval! Thus, knowledge of the spectrum for one interval yields knowledge of it for all intervals and hence q is constructed via a trace formula.

The main purpose of the present work is to obtain asymptotic and trace formulas for more general operators L_0 and $L = L_0 + q$. The formulas we derive reveal certain interesting quantitative and asymptotic properties of the spectrum of a variety of eigenvalue problems that appear quite often in mathematical physics, engineering, etc. In particular, the formulas presented here can be helpful in solving inverse problems.

Finally, there seems to be a theoretical application of the results of this work, namely, since in some sense $L - L_0 = q$, the trace formulas suggest a concept of trace for multiplication operators $(Tf)(x) = q(x)f(x)$, when q is smooth and satisfy certain zero-average conditions.

1. Preliminaries. Let L_0 and L be the operators on $L^2(0,b)$ defined by

$$L_0 u = -u'' \quad \text{and} \quad Lu = L_0 u + q(x)u(x), \tag{1.1}$$

where $q \in C^2[0,b]$ and $0 < b < \infty$. We assume Dirichlet (i.e., zero) boundary conditions for the functions in the domain of L_0 (and L), in order to have a unique well-defined self-adjoint operator. We call $\mu_1 < \mu_2 < \cdots < \mu_n < \ldots$ the eigenvalues of L and $\phi_1(x), \phi_2(x), \ldots, \phi_n(x), \ldots$, the associated orthonormal eigenfunctions. Similarly, $\nu_1 < \nu_2 < \cdots < \nu_n < \ldots$ and $\psi_1(x), \psi_2(x), \ldots, \psi_n(x), \ldots$, are the corresponding quantities for L_0. In fact, for this particular case we have

$$\nu_n = \frac{\pi^2 n^2}{b^2} \qquad \text{and} \qquad \psi_n(x) = \sqrt{\frac{2}{b}} \sin \frac{\pi n x}{b}. \tag{1.2}$$

In the early 1950s Levitan and others proved that

$$\mu_n = \nu_n + \frac{1}{b}\int_0^b q(x)dx + O(n^{-2}). \tag{1.3}$$

Their proof can be found in [L-G]. A simpler proof was given later by Hochstadt with the help of an ingenious transformation (see [H.H]). These methods have the advantage of giving the full asymptotic behaviour of μ_n, but the big disadvantage is that they work only for these particular operators.

From now on (and without loss of generality) we will assume that the average of q on $[0,b]$ is zero, namely

$$\int_0^b q(x)dx = 0 \tag{1.4}$$

(since the average of q causes only a constant shift of the spectrum).

A simple, general, and convincing (but not rigorous) way of proving (1.3) is by using the ideas of the WKB approximation (from Wentzel, Kramers, and Brillouin—this approach was shown to us by Professor Venakides). We discuss this very interesting method in the next section.

Yet there is a third method of obtaining a result of the type of (1.3). It can be applied to much more general situations, it is rigorous and abstract, but it yields weaker results. It is based on the following lemma.

LEMMA. *Let A be a self-adjoint operator (bounded or unbounded) acting on $L^2(D)$, where D is a domain in R^d. Consider the family of operators*

$$H(s) = A + sq(x), \qquad s \in [0,1]. \tag{1.5}$$

For convenience, let us assume that q is in $C^r(\bar{D})$, where we can take r as big as we wish.

We also assume that, for each s, the operator $H(s)$ has a discrete spectrum

$$\lambda_1(s) < \lambda_2(s) < \cdots < \lambda_n(s) < \ldots.$$

(Notice that, given the fact that $\lambda_n(s)$ is continuous in s, the assumption implies that the eigenvalues of $H(s)$, being simple for all s, never cross each other. It is, therefore, a very strong assumption, especially in higher dimensions.)

Then

$$\frac{d\lambda_n}{ds} = (q\chi_n, \chi_n) \tag{1.6}$$

and

$$\frac{d^2\lambda_n}{ds^2} = -2(q\chi_n, R_n(s)q\chi_n), \tag{1.7}$$

where $(\cdot\,,\cdot)$ is the inner product of $L^2(D)$, $\chi_n(x;s)$ is the eigenfunction of $H(s)$ that corresponds to $\lambda_n(s)$, normalized such that $(\chi_n, \chi_n) = 1$, and $R_n(s) = [I - P_n(s)][H(s) - \lambda_n(s)]|-1$, $P_n(s)$ being the (orthogonal) projection onto χ_n. (The proof can be found in [R-S, §13.16].

Now, the standard Taylor's theorem with remainder gives

$$|\lambda_n(1) - \lambda_n(0) - \lambda_n'(0)| \le \tfrac{1}{2} \sup_{0\le s\le 1} |\lambda_n''(s)|. \tag{1.8}$$

In certain cases (for example, see §4), (1.7) can help us to obtain a good upper bound for the right-hand side of (1.8). For instance, if A is L_0 of (1.1) and q is in $C^1[0,b]$, we get easily that this upper bound is $O(n^{-1})$, which is a weaker version of (1.3), but requires less smoothness for q (notice that the Riemann–Lebesgue lemma is involved). For the details see [R-S, Thm. 13.82.5].

Recently, Friedlander (see [F]) and Feldman, Knorrer, and Trubowitz (see [F-K-T]) obtained some results concerning the distribution of the Floquet eigenvalues of the operator $-\Delta + q$, where q is periodic on $\mathbf{R}^2$. Their findings remind (1.3), but there are exceptional zero-density sets of eigenvalues which do not obey the asymptotics.

Few years after the discovery of (1.3), Gelfand (see [G-L]) observed that, assuming (1.4), i.e., zero average for q, one has the beautiful trace formula

$$\operatorname{tr}(L - L_0) \overset{\text{def}}{=} \sum_n \left[\mu_n - \frac{\pi^2 n^2}{b^2}\right] = \frac{q(0) + q(b)}{4}. \tag{1.9}$$

The following proof of (1.9) appeared in a lecture given by Douady (see [D-H-V]). It utilizes the heat kernel in a neat way and it inspired the derivations of the other trace formulas in the present work. This is why we decided to include it in our introduction.

Let L and L_0 be as in (1.1) and, for $t > 0$, let $k(t,x,y)$ and $p_b(t,x,y)$ be the integral kernels of the (compact) operators e^{-tL} and e^{-tL_0}, respectively. In other words, $k(t,x,y)$ and $p_b(t,x,y)$ are the heat kernels of L and L_0. If we set

$$p_\infty(t,x,y) = \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{|x-y|^2}{4t}\right) - \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{|x+y|^2}{4t}\right) \tag{1.10a}$$

(which, incidentally, is the heat kernel of L_0 acting on $(0,\infty)$, with Dirichlet boundary condition at 0), then it can be easily checked that (see, for example, [J])

$$p_b(t,x,y) = \sum_{k\in Z} p_\infty(t,x,y+2kb), \qquad 0 \le x, \quad y \le b. \tag{1.10b}$$

Thus, as $t \downarrow 0$, we have uniformly in $x, y \in [0,b]$ that

$$p_b(t,x,y) = \frac{1}{\sqrt{4\pi t}}\left[\exp\left(-\frac{|x-y|^2}{4t}\right) - \exp\left(-\frac{|x+y|^2}{4t}\right) - \exp\left(-\frac{|x+y-2b|^2}{4t}\right)\right] + \frac{e^{-c/t}}{\sqrt{t}} O(1),$$

where $c = b^2/4$. In particular,

$$p_b(t,x,x) = \frac{1}{\sqrt{4\pi t}}\left[1 - \exp\left(-\frac{x^2}{t}\right) - \exp\left(-\frac{|x-b|^2}{t}\right)\right] + \frac{e^{-c/t}}{\sqrt{t}} O(1). \tag{1.11}$$

Also, it is not hard to obtain the following perturbation expansion for $k(t,x,y)$

$$k(t,x,y) = \sum_{n=0}^{\infty} k_n(t,x,y), \tag{1.12a}$$

where

$$k_0(t,x,y) = p_b(t,x,y) \tag{1.12b}$$

and, for $n \ge 1$,

$$k_n(t,x,y) = (-1)^n \int_0^t \int_0^b p_b(s,x,z) q(z) k_{n-1}(t-s,z,y)\, dz\, ds. \tag{1.12c}$$

Equation (1.12c) implies

$$|k_n(t,x,y)| \le \frac{1}{n!} \|q\|_\infty^n t^n p_b(t,x,y). \tag{1.13}$$

From (1.13) we get immediately

$$k_n(t,x,y) = O(t^{n-1/2}), \quad \text{as } t \downarrow 0, \text{ uniformly in } x, y. \tag{1.14}$$

Next we write the eigenfunction expansions of $k(t,x,y)$ and $p_b(t,x,y)$

$$k(t,x,y) = \sum_n e^{-\mu_n t}\phi_n(x)\phi_n(y),$$

$$p_b(t,x,y) = \sum_n e^{-\nu_n t}\psi_n(x)\psi_n(y).$$

Thus, by the orthonormality

$$\int_0^b [k(t,x,x) - p_b(t,x,x)]dx = \sum_n [e^{-\mu_n t} - e^{-\nu_n t}] = \mathrm{tr}(e^{-tL} - e^{-tL_0}).$$

We need the short-time asymptotics because, at least formally we have

$$\mathrm{tr}(L - L_0) = \lim_{t\downarrow 0} \frac{1}{t}\mathrm{tr}(e^{-tL} - e^{-tL_0}).$$

Using formulas (1.12) and (1.14) we get easily (as $t \downarrow 0$)

$$\int_0^b k_1(t,x,x)dx = \sum_n [e^{-\mu_n t} - e^{-\nu_n t}] + O(t^{3/2}). \tag{1.15}$$

But by (1.12) and Fubini–Tonelli,

$$\begin{aligned}
\int_0^b k_1(t,x,x)dx &= -\int_0^b \int_0^t \int_0^b p_b(s,x,z)q(z)p_b(t-s,z,x)dzdsdx \\
&= -\int_0^b q(z)\left[\int_0^t \int_0^b p_b(s,x,z)p_b(t-s,z,x)dsdx\right]dz \\
&= -\int_0^b q(z)\left[\int_0^t p_b(t,z,z)ds\right]dz \\
&= -t\int_0^b p_b(t,z,z)q(z)dz
\end{aligned}$$

and thus, because of (1.11)

$$\int_0^b k_1(t,x,x)dx = -\frac{\sqrt{t}}{\sqrt{4\pi}}\int_0^b q(x)dx + \frac{q(0)}{4}t + \frac{q(b)}{4}t + O(t^{3/2}), \tag{1.16}$$

since, if q is in $C^1[0,b]$,

$$\frac{1}{\sqrt{4\pi t}}\int_0^b \exp\left(-\frac{x^2}{t}\right)q(x)dx = \frac{q(0)}{4} + O(\sqrt{t}) \quad \text{as } t \downarrow 0. \tag{1.17}$$

If we assume (1.4), namely

$$\int_0^b q(x)dx = 0,$$

and combine (1.15) and (1.16) we arrive at (1.9) by dividing by t and letting $t \downarrow 0$. We need (1.3) in order to interchange summation and limit, since (1.3) guarantees that $(\mu_n - \nu_n)$ is absolutely summable.

In what follows, we attempt to extend the formulas and ideas presented above.

In §2 we give the vector analog of (1.3) and (1.16). To be more precise, we analyze the case $L_0u = (-d^2/dx^2)u$ and $Lu = L_0u + Q(x)u$, where $Q(x)$ is an $r \times r$ real symmetric matrix and $u = (u_1, \dots, u_r)$ with $u_j \in L^2(0,b)$. We denote this by $u \in L_r^2(0,b)$. This is a Hilbert space and L has a unique self-adjoint extension on $L_r|2(0,b)$ if we prescribe boundary conditions, say, $u(0) = u(b) = 0$. The importance of this problem lies on the idea that, by carefully choosing Q and then letting $r \to \infty$, we might obtain interesting results for the two-dimensional operator $-\Delta + q$.

In §3 we examine the case where $L_0 = -\Delta$ and $L = L_0 + q$ are acting on $L^2(D)$ and where $D = (0,b_1) \times (0,b_2)$. We get two trace formulas, but they cannot reduce to something completely analogous to (1.16) because, to our knowledge, in the two-dimensional case there is no asymptotic formula as strong as (1.3).

In §4 we let $L_0 = (-d^2/dx^2)^m$ and $L = L_0 + q$, acting on $L^2(0,b)$. Here we are able to obtain the complete analogs of (1.3) and (1.16). Formula (4.15) of this section is not new. It exists in the Russian literature (see, for example, [L-S]), but our way of deriving it is different and, in our opinion, simpler.

In §5 we extend the results of §4 to the two-dimensional case. Many of the estimates needed here have been developed in the previous section.

The last section contains some final thoughts. In particular we make some remarks for the case where $L_0 = -\Delta$ with Neumann boundary conditions and $L = -\Delta$ with boundary condition

$$\frac{\partial u}{\partial n}(z) + c(z)u(z) = 0, \qquad z \in \partial D.$$

Here the trace formula involves the boundary function $c(z)$.

2. The vector-valued function case. We define

$$L_r^2(0,b) = \{u = (u_1, \dots, u_r) : u_j \in L^2(0,b),\ 1 \le j \le r\}.$$

This is a Hilbert space with inner product

$$(u,v) = \int_0^b u \cdot v\, dx = \int_0^b (u_1v_1 + \cdots + u_rv_r)dx,$$

where $u = (u_1, \dots, u_r)$ and $v = (v_1, \dots, v_r)$.

For a sufficiently smooth u in $L_r^2(0,b)$ we set

$$L_0u = -\frac{d^2u}{dx^2} + Q_0u \quad \text{and} \quad Lu = L_0u + Q(x)u, \tag{2.1}$$

where Q_0 is an $r \times r$ symmetric matrix with real and constant entries and $Q(x) = [q_{ij}(x)]_{1 \le i,j \le r}$ is also real, symmetric, and such that each q_{ij} is in $C^2[0,b]$ and

$$\int_0^b q_{ij}(x)dx = 0. \tag{2.2}$$

We consider the boundary conditions

$$u(0) = u(b) = 0. \tag{2.3}$$

Then L is a well-defined self-adjoint operator on $L_r^2(0,b)$ and, without loss of generality, Q_0 is a diagonal matrix, namely

$$Q_0 = \operatorname{diag}(a_1, a_2, \dots, a_r), \qquad a_1 \le a_2 \le \cdots \le a_r \tag{2.4}$$

(since Q_0 can be diagonalized by an orthogonal transformation independent of x). The spectrum of L is discrete and has ∞ as its only cluster point (since the variation of parameters method implies that $(L-z)^{-1}$ is compact, etc).

Let us denote by $\mu_1, \mu_2, \ldots, \mu_n, \ldots$ the spectrum of L and by $\nu_1, \nu_2, \ldots, \nu_n, \ldots$ the spectrum of L_0 as usual. Here is the analog of (1.3).

CONJECTURE.

$$\mu_n = \nu_n + O(n^{-2}). \tag{2.5}$$

Proof (adaptation of the WKB method, shown to us by Stephanos Venakides). For convenience we assume (essentialy without loss of generality) that the diagonal entries of Q_0 are distinct. Then, the eigenvalues of L_0 (which decouples to r very simple scalar operators) are

$$\nu_{nk} = a_k + \frac{\pi^2 n^2}{b^2}, \qquad 1 \le k \le r,$$

with corresponding eigenfunctions

$$\psi_{nk}(x) = e_k \sin \frac{\pi n x}{b},$$

where $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, i.e., all its entries are zero except for the kth one, which is 1. The assumption that the a_k's are distinct makes the eigenvalues of L_0 simple, at least for all n sufficiently large.

Now we view L as a perturbation of L_0 and thus we make the (unjustified) assumption that for large n, its eigenvalues $\{\mu_{nk}\,;\, 1 \le k \le r,\ n = 1, 2, \ldots\}$ and eigenfunctions $\{\phi_{nk}(x)\,;\, 1 \le k \le r,\ n = 1, 2, \ldots\}$ can be written as expansions in powers of n^{-1}, namely

$$L\phi_{nk} = (\nu_{nk} + c_{-1k}n + c_{0k} + c_{1k}n^{-1} + c_{2k}n^{-2} + \cdots)\phi_{nk} \tag{2.6}$$

and

$$\phi_{nkj}(x) = [A_{0kj}(x) + A_{1kj}(x)n^{-1} + \cdots] \sin[\pi n x/b + \omega_{0kj}(x) + \omega_{1kj}(x)n^{-1} + \cdots],$$

where ϕ_{nkj} is the jth coordinate of ϕ_{nk}. Substituting this in (2.6) and equating coefficients of the same powers of n, we obtain after some algebra that $c_{-1k} = c_{0k} = c_{1k} = 0$, which is (2.5). Formula (2.2) and the boundary conditions play an essential role, as expected. □

We continue with the trace formula. The integral kernel of the semigroup e^{-tL} is given by a formula similar to (1.12), whose probabilistic version (to be a little fancy) is the Feynman–Kac formula

$$k(t, x, y) = E^x \left\{ \exp\left[-tQ_0 - \frac{1}{2}\int_0^{2t} Q(X_s)ds \right] \,\middle|\, X_{2t} = y \right\} p_b(t, x, y), \tag{2.7}$$

where X_s is the Brownian motion process in $(0, b)$ killed at 0 and b and $p_b(t, x, y)$ is its transition density, given by (1.10b). The eigenfunction expansion of this heat kernel is

$$k(t, x, y) = \sum_n e^{-\mu_n t}\phi_n(x) \otimes \phi_n(y),$$

where $\phi_n(x) \otimes \phi_n(y)$ is the $r \times r$ matrix whose ijth entry is $\phi_{ni}(x)\phi_{nj}(y)$, ϕ_{ni} being the ith component of the (normalized) nth eigenfunction ϕ_n of L. A simple step by step adaptation of the proof of (1.9) gives

$$\lim_{t \downarrow 0} \frac{1}{t} \sum_n (e^{-t\mu_n} - e^{-t\nu_n}) = \frac{\mathrm{tr}Q(0) + \mathrm{tr}Q(b)}{4} \tag{2.8}$$

and thus, if (2.5) is true, we can interchange summation and limit and get

$$\mathrm{tr}(L - L_0) = \sum_n (\mu_n - \nu_n) = \frac{\mathrm{tr}Q(0) + \mathrm{tr}Q(b)}{4}. \tag{2.8'}$$

Remark. One good reason for studying operators of the type of L is that, by choosing Q_0 and Q in a suitable way and then letting $r \to \infty$, we might be able to obtain spectral properties of higher-dimensional operators (for example, $-\Delta + q$).

3. The case of a rectangle. Consider the rectangular domain $D = (0, b_1) \times (0, b_2)$ in R^2, where $0 < b_1, b_2 < \infty$. Let L_0 and L be the operators on $L^2(D)$ defined by

$$L_0 u = -\Delta u \quad \text{and} \quad Lu = L_0 u + q(x)u(x), \tag{3.1}$$

where $q \in C^2(\bar{D})$. We assume again Dirichlet (i.e., zero) boundary conditions for the functions in the domain of L_0 (and L), in order to have a unique well-defined self-adjoint operator. We call, as in the introduction, $\mu_1 < \mu_2 \leq \cdots \leq \mu_n \leq \cdots$ the eigenvalues of L and $\phi_1(x), \phi_2(x), \ldots, \phi_n(x), \ldots$ the associated orthonormal eigenfunctions. Similarly, $\nu_1 < \nu_2 \leq \cdots \leq \nu_n \leq \ldots$ and $\psi_1(x), \psi_2(x), \ldots, \psi_n(x), \ldots$ are the corresponding quantities for L_0. Here we have

$$\nu_n = \frac{\pi^2 n_1^2}{b_1^2} + \frac{\pi^2 n_2^2}{b_2^2} \quad \text{and} \quad \psi_n(x) = \frac{2}{\sqrt{b_1 b_2}} \sin \frac{\pi n_1 x_1}{b_1} \sin \frac{\pi n_2 x_2}{b_2}, \tag{3.2a}$$

where n_1 and n_2 are integers depending on n so that we always have $\nu_n \leq \nu_{n+1}$. The asymptotic behaviour of ν_n for large n is (see [C-H], vol. 1, Chap. 6, §4)

$$\nu_n = \frac{4\pi}{b_1 b_2} n + O(\sqrt{n}). \tag{3.2b}$$

The estimate

$$|\mu_n - \nu_n| \leq \|q\|_\infty \tag{3.3}$$

follows easily from the standard minimax argument. Stephanos Venakides asked whether we can say something better than (3.3). We suspect that there must be a better estimate (except for a set of eigenvalues of zero density) when the averages of q along the x_1 and x_2 axes are zero. This is suggested by the formulas (3.12) and (3.16) and by the recent works [F] and [F-K-T] for the case of periodic boundary conditions.

Let us imitate the analysis at the end of §1. The heat kernel $p(t, x, y)$ of L_0 is given by (separation of variables)

$$p(t, x, y) = p_1(t, x_1, y_1)p_2(t, x_2, y_2), \tag{3.4}$$

where, for $j = 1$ or 2, $p_j(t,x,y) = p_{b_j}(t,x,y)$ of (1.17). The heat kernel $k(t,x,y)$ of L is

$$k(t,x,y) = \sum_{n=0}^{\infty} k_n(t,x,y), \tag{3.5a}$$

where

$$k_0(t,x,y) = p(t,x,y) \tag{3.5b}$$

and, for $n \geq 1$,

$$k_n(t,x,y) = -\int_0^t \int_D p(s,x,z)q(z)k_{n-1}(t-s,z,y)dzds. \tag{3.5c}$$

Estimate (1.22) becomes

$$|k_n(t,x,y)| \leq \frac{1}{n!}\|q\|_\infty^n t^n p(t,x,y). \tag{3.6}$$

Thus

$$k_n(t,x,y) = O(t^{n-1}), \quad \text{as } t \downarrow 0, \text{ uniformly in } x,y. \tag{3.7}$$

The eigenfunction expansions give (as in §1)

$$\int_D [k(t,x,x) - p_b(t,x,x)]dx = \sum_n (e^{-\mu_n t} - e^{-\nu_n t}).$$

Because of (3.5) and (3.6) this implies

$$\int_D [k_1(t,x,x) + k_2(t,x,x)]dx = \sum_n (e^{-\mu_n t} - e^{-\nu_n t}) + O(t^2), \quad \text{as } t \downarrow 0. \tag{3.8}$$

Now the same argument used in §1 for the one-dimensional case gives, as $t \downarrow 0$,

$$\int_D k_1(t,x,x)dx = -t\int_D p(t,x,x)q(x)dx,$$

thus, because of (3.4) and (1.17),

(3.9a)

$$\begin{aligned}\int_D k_1(t,x,x)dx &= -t\int_0^{b_2} p_2(t,x_2,x_2)\left[\int_0^{b_1} p_1(t,x_1,x_1)q(x_1,x_2)dx_1\right]dx_2 \\ &= -\frac{b_1 b_2 Q}{4\pi} + \frac{(b_1+b_2)Q}{\sqrt{4\pi}}\sqrt{t} + \frac{b_1[Q_2(0)+Q_2(b_2)] + b_2[Q_1(0)+Q_1(b_1)]}{4\sqrt{4\pi}}\sqrt{t} \\ &\quad - \frac{q(0,0)+q(b_1,0)+q(0,b_2)+q(b_1,b_2)}{16}t + O(t\sqrt{t}),\end{aligned}$$

where

$$\begin{aligned} Q &= \frac{1}{b_1 b_2}\int_D q(x)dx, \quad Q_1(x_1) = \frac{1}{b_2}\int_0^{b_2} q(x_1,x_2)dx_2, \\ Q_2(x_2) &= \frac{1}{b_1}\int_0^{b_1} q(x_1,x_2)dx_1. \end{aligned} \tag{3.10}$$

Also, from (1.12) and (1.13)

$$\int_D k_2(t,x,x)dx = \frac{t}{8\pi}\int_D q(x)^2dx + O(t\sqrt{t}), \qquad \text{as } t \downarrow 0, \tag{3.9b}$$

thus, by substituting in (3.8) we arrive at the following.

THEOREM. *As* $t \downarrow 0$,

$$\begin{aligned}\sum_n (e^{-\mu_n t} - e^{-\nu_n t}) &= -\frac{b_1 b_2 Q}{4\pi} + \frac{(b_1+b_2)Q}{\sqrt{4\pi}}\sqrt{t} + \frac{b_1[Q_2(0)+Q_2(b_2)] + b_2[Q_1(0)+Q_1(b_1)]}{4\sqrt{4\pi}}\sqrt{t} \\ &\quad - \frac{q(0,0)+q(b_1,0)+q(0,b_2)+q(b_1,b_2)}{16}t + \frac{t}{8\pi}\int_D q(x)^2dx + O(t\sqrt{t}),\end{aligned}$$

where the Qs are given by (3.10).

Remark. Suppose that $q(x_1+b_1,x_2) = q(x_1,x_2+b_2) = q(x_1,x_2)$, i.e., q is periodic with period D. Then the theorem implies that q can be recovered from the family of Dirichlet spectra $\{\mu_n(\xi)\}_{n=1}^{\infty}$ of L_ξ, $\xi \in D$, where

$$L_\xi u(x) = -\Delta u(x) + q(x+\xi)u(x).$$

In the one-dimensional case there are "evolution" equations (in ξ) for the $\mu_n(\xi)$'s. Here this does not seem to be the case.

If in (3.10)

$$Q_1(x_1) \equiv Q_2(x_2) \equiv Q = 0 \tag{3.11}$$

(which is not really a strong assumption on q, since $q(x_1,x_2) = \tilde{q}(x_1,x_2) + Q_1(x_1) + Q_2(x_2) - Q$, where $\tilde{q}$ satisfies this assumption or, if we consider the Fourier expansion of q in D, the assumption means that there are no terms in this double series which depend only on x_1 or only on x_2), then the theorem implies

$$\lim_{t\downarrow 0}\frac{1}{t}\sum_n (e^{-\mu_n t} - e^{-\nu_n t}) = -\frac{q(0,0)+q(b_1,0)+q(0,b_2)+q(b_1,b_2)}{16} + \frac{1}{8\pi}\int_D q(x)^2dx, \tag{3.12}$$

but here we cannot pass the limit inside the sum in order to get the complete analog of (1.9), because the asymptotics of $\mu_n - \nu_n$ look messier than in the one-dimensional case, although formula (3.12) suggests (see also [F] and [F-K-T]) that there we can expect better than (3.3).

Before closing this section we want to make a final observation. Assume that (3.11) holds and, for convenience, let the spectrum of L be strictly positive, namely let

$$\mu_1 > 0$$

(otherwise we just need to multiply everything below by e^{-tc}, where c is a constant such that $c + \mu_1 > 0$). Then, if we set

$$\tilde{k}(t,x,y) = k(t,x,y) - p(t,x,y),$$

we have the long-time estimate

$$\tilde{k}(t,x,y) = O(e^{-t(\mu_1 \wedge \nu_1)}), \quad \text{as } t \to \infty, \quad \text{uniformly in } x, y. \tag{3.13}$$

Of course, as we have already seen

$$\sum_n (e^{-\mu_n t} - e^{-\nu_n t}) = \int_D \tilde{k}(t,x,x)dx.$$

Let $\beta > 0$. If we replace t by t^β in the above formula, then divide by t^β and finally integrate with respect to t from 0 to ∞, we get

$$\int_0^\infty \frac{1}{t^\beta}\left[\sum_n (e^{-\mu_n t^\beta} - e^{-\nu_n t^\beta})\right] dt = \int_0^\infty \frac{1}{t^\beta}\left[\int_D \tilde{k}(t^\beta,x,x)dx\right] dt. \tag{3.14}$$

The right-hand side of the above equation is well defined (i.e., converges) for all $\beta > 0$ because of (3.12) and (3.13). To understand a little better the left-hand side, observe that, for $a, b > 0$,

$$\int_0^\infty \frac{e^{-at^\beta} - e^{-bt^\beta}}{t^\beta} dt = \begin{cases} \frac{\Gamma(1/\beta)}{\beta - 1}[b^{1-1/\beta} - a^{1-1/\beta}], & \text{if } \beta \neq 1; \\ \ln b - \ln a, & \text{if } \beta = 1. \end{cases} \tag{3.15}$$

Next we notice that (3.2b) and (3.3) imply (even without (3.11))

$$\sum_n |\mu_n^{-\epsilon} - \nu_n^{-\epsilon}| < \infty \quad \text{for every } \epsilon > 0$$

and therefore, if $0 < \beta < 1$, we can interchange summation and integral in the left-hand side of (3.14) and then use (3.15a) to obtain

$$\begin{aligned} \operatorname{tr}(L^{1-1/\beta} - L_0^{1-1/\beta}) &= \frac{\Gamma(1/\beta)}{\beta - 1} \sum_n [\nu_n^{1-1/\beta} - \mu_n^{1-1/\beta}] \\ &= \int_0^\infty \frac{1}{t^\beta}\left[\int_D \tilde{k}(t^\beta,x,x)dx\right] dt. \end{aligned} \tag{3.16}$$

The question is whether (3.16) is true for any $\beta \geq 1$. The hope that some result of this kind exists comes from the fact that (as we have already pointed out) the right-hand side of (3.16) makes sense for all $\beta > 0$. In fact we can even let $\beta \to \infty$ in the right-hand side of (3.16)

$$\begin{aligned} &\lim_{\beta\to\infty} \int_0^\infty \frac{1}{t^\beta}\left[\int_D \tilde{k}(t^\beta,x,x)dx\right] dt \\ &\quad = \lim_{\beta\to\infty} \int_0^1 \frac{1}{t^\beta}\left[\int_D \tilde{k}(t^\beta,x,x)dx\right] dt + \lim_{\beta\to\infty} \int_1^\infty \frac{1}{t^\beta}\left[\int_D \tilde{k}(t^\beta,x,x)dx\right] dt \\ &\quad = -\frac{q(0,0) + q(b_1,0) + q(0,b_2) + q(b_1,b_2)}{16} + \frac{1}{8\pi}\int_D q(x)^2 dx, \end{aligned}$$

where the first limit is computed by (3.12) and dominated convergence whereas the second limit is 0 by (3.13).

4. Higher-order Schrödinger operators I. We call the operator

$$L = (-\Delta)^m + q \tag{4.1}$$

an *m*th-*order Schrödinger operator.* Such operators appear in many engineering models, for example, in the study of vibrating plates or hydraulic flow, to name a few.

To construct the heat kernel of L (namely the integral kernel of the operator e^{-tL}), we first consider the projection operators E_λ associated to the Laplacian $-\Delta$, acting on $L^2(R^d)$. It is well known (see [S]) that these projections are integral operators. Let $e_d(x, y; \lambda)$ be the integral kernel of E_λ. Since

$$\int_0^\infty e^{-t\lambda} d_\lambda e_d(x, y; \lambda) = \frac{1}{(4\pi t)^{d/2}} \exp\left(-\frac{|x-y|^2}{4t}\right), \tag{4.2}$$

the tables for inverse Laplace transforms give

(4.3)

$$e_d(x, y; \lambda) = \frac{\lambda^{d/4}}{(2\pi|x-y|)^{d/2}} J_{d/2}\big(|x-y|\sqrt{\lambda}\big), \qquad e_d(x, x; \lambda) = \frac{\lambda^{d/2}}{(4\pi)^{d/2}\Gamma[(d/2)+1]},$$

where

$$J_{d/2}(z) = (z/2)^{d/2} \sum_{k=0}^{\infty} \frac{(-z^2/4)^k}{k!\Gamma[k + (d/2) + 1]}$$

is the Bessel function of order $d/2$. Notice that, if d is odd, $J_{d/2}(z)$ is an elementary function. For example,

$$J_{1/2}(z) = \sqrt{\frac{2}{\pi}} \frac{\sin z}{\sqrt{z}}$$

and therefore

$$e_1(x, y; \lambda) = \frac{1}{\pi} \frac{\sin\big(|x-y|\sqrt{\lambda}\big)}{|x-y|}. \tag{4.3′}$$

We also notice that, for any fixed $\lambda > 0$ (and d), $e_d(x, y; \lambda) = f(|x-y|)$, where f is an even entire function of order $1/2$.

In the spirit of the previous sections we set

$$L_0 = (-\Delta)^m. \tag{4.1′}$$

The heat kernel of L_0 in $R|d$ is given by

$$H_d^m(t, x, y) = h_d^m(t, |x-y|) = \int_0^\infty e^{-t\lambda^m} d_\lambda e_d(x, y; \lambda), \tag{4.4}$$

but now $H_d^m(t, x, y) \neq H_1^m(t, x_1, y_1) \dots H_1^m(t, x_d, y_d)$, if $m > 1$. Observe that $h_d^m(t, r)$ is even in r. From the asymptotics of the Bessel functions for a large argument we can conclude that, if $x \neq y$ then $H_d^m(t, x, y) \to 0$ as $t \downarrow 0$, but, if $x = y$, then (4.3) gives

$$\begin{aligned} H_d^m(t, x, x) &= \frac{1}{(4\pi)^{d/2}\Gamma(d/2)} \int_0^\infty e^{-t\lambda^m} \lambda^{(d-2)/2} d\lambda \\ &= \frac{\Gamma(d/2m)}{m(4\pi)^{d/2}\Gamma(d/2)} \frac{1}{t^{(d/2m)}} = C_d^m \frac{1}{t^{(d/2m)}}, \end{aligned} \tag{4.4′}$$

which blows up when $t \downarrow 0$ as expected, since it is the kernel of a semigroup.

Remark. Equations (4.4) and (4.3) imply that

$$H_d^m(t,x,y) = \frac{1}{t^{d/2m}} f_d^m\left(\frac{|x-y|}{t^{1/2m}}\right),$$

where $f_d^m(r)$ is even, entire and, for real r, it is in the Schwartz class, as it follows from Lemma 4.1 below. Therefore,

$$\int_{R^d} |H_d^m(t,x,y)|dy = \int_{R^d} \left| f_d^m\left(\sqrt{y_1^2 + \cdots + y_d^2}\right)\right| dy < \infty.$$

The fact that H_d^m is a heat kernel implies that f_d^m must satisfy the equation

$$(-1)^m \Delta^m f_d^m(r) = \frac{r}{2m}\frac{d}{dr} f_d^m(r) + \frac{d}{2m} f_d^m(r),$$

Δ being the radial Laplacian in R^d, i.e.,

$$\Delta u(r) = u''(r) + \frac{d-1}{r} u'(r).$$

Furthermore, since $f_d^m(r) = H_d^m(1,0,y)$, with $r = |y|$, we obtain from (4.4) (and (4.6)) the integral representation (which is in fact a Hankel transorm)

$$f_d^m(r) = \frac{1}{(2\pi)^{d/2} r^{(d-1)/2}} \int_0^\infty e^{-x^{2m}} x^{(d-1)/2} J_{(d-2)/2}(rx)\sqrt{rx}\,dx, \quad d \geq 2$$

and

$$f_1^m(r) = \frac{1}{\pi}\int_0^\infty e^{-x^{2m}} \cos(rx)dx.$$

Using steepest descents in the integral representation for f_d^m (see [B-O, Chap. 6, §6]) or, better, the method of dominant balance (see [B-O, Chap. 3, §4]—notice that f_d^m is in the Schwartz class) in the differential equation that f_d^m satisfies, we obtain

$$f_d^m(r) \sim K \exp\left(-r^{2m/2m-1} c\beta\right) \cos\left(r^{2m/2m-1} c\alpha\right), \quad \text{as } r \to \infty,$$

where K is a constant and

$$c = \frac{d(2m-1)}{(2m)^{2m/2m-1}}, \quad \alpha = \cos\left(\frac{\pi}{4m-2}\right), \quad \beta = \sin\left(\frac{\pi}{4m-2}\right).$$

Observe that, unless $m = 1$, $f_d^m(r)$ takes both positive and negative values because of the appearance of the cosine. Thus $H_d^m(t,x,y)$ cannot define a probability transition density, but it still defines a (finite) measure on the functions with domain $[0,\infty)$ and range R^d, similar to the Wiener measure (except for the nonnegativity and the almost sure continuity of the paths). For more details, especially for the case $d = 1$, see [H].

To continue we need the asymptotics of $h_d^m(t,r)$ of (4.4), as $r = |x-y| \to \infty$.

LEMMA 4.1. *As* $r \to \infty$

$$\frac{d^l}{dr^l} h_d^m(t,r) = o(r^{-k}), \quad \textit{for any } k > 0,\ l \geq 0 \tag{4.5}$$

(*i.e.*, $h_d^m(t,\cdot)$ *is in the Schwartz class* $\mathcal{S}(R)$).

Proof. We define

$$g_1(\lambda) = e_d(x,y;\lambda), \qquad g_{n+1}(\lambda) = \int_0^\lambda g_n(\xi)d\xi, \quad n \geq 1.$$

Thus, by applying simple integration by parts in (4.2) we obtain

$$\int_0^\infty e^{-t\lambda} g_n(\lambda)d\lambda = \frac{1}{t^n(4\pi t)^{d/2}} \exp\left(-\frac{|x-y|^2}{4t}\right),$$

hence the tables for inverse Laplace transforms imply

$$g_n(\lambda) = \frac{2^{n-1}}{(2\pi)^{d/2}} \frac{\lambda^{\nu/2}}{|x-y|^\nu} J_\nu\big(|x-y|\sqrt{\lambda}\big),$$

where $\nu = n + (d/2) - 1$. Therefore, if we apply integration by parts in the integral of (4.4) n times, we will get

$$h_d^m(t,r) = \frac{C_n}{r^\nu} \int_0^\infty e^{-t\lambda^m} p(\lambda) J_\nu\big(r\sqrt{\lambda}\big)d\lambda,$$

where C_n is a constant that depends only on n, $p(\lambda)$ grows polynomially in λ, and ν is as above. □

Remark. We also have from (4.2) that

$$\frac{de_d}{d\lambda}(x,y;\lambda) = \frac{1}{2(2\pi)^{d/2}} \frac{\lambda^{(d-2)/4}}{|x-y|^{(d-2)/2}} J_{(d-2)/2}\big(|x-y|\sqrt{\lambda}\big), \quad \text{if } d \geq 2 \tag{4.6}$$

and, if $d = 1$,

$$\frac{de_1}{d\lambda}(x,y;\lambda) = \frac{1}{2\pi} \frac{\cos\big(|x-y|\sqrt{\lambda}\big)}{\sqrt{\lambda}}. \tag{4.6′}$$

For the remaining of this section we will restrict ourselves to the case $d = 1$.

The heat kernel of L_0 in $L^2(0,b)$ with ("generalized") Dirichlet boundary conditions (namely, $u(0) = u^{(2)}(0) = \cdots = u^{2(m-1)}(0) = 0$ and $u(b) = u^{(2)}(b) = \cdots = u^{2(m-1)}(b) = 0$), can be constructed from $h_1^m(t,|x-y|)$ by the method of images as in (1.10), thanks to the previous lemma.

$$p_b^m(t,x,y) = \sum_{k\in Z} p_\infty^m(t,x,y+2kb), \qquad 0 \leq x,y \leq b, \tag{4.7a}$$

where from (4.4) and (4.6′) we get

$$\begin{aligned} h_1^m(t,|x-y|) &= \frac{1}{2\pi} \int_0^\infty e^{-t\lambda^m} \frac{\cos\big(|x-y|\sqrt{\lambda}\big)}{\sqrt{\lambda}} d\lambda \\ &= \frac{1}{\pi} \int_0^\infty e^{-ts^{2m}} \cos\big(|x-y|s\big)ds \end{aligned} \tag{4.7b}$$

and

$$p_\infty^m(t,x,y) = h_1^m(t,|x-y|) - h_1^m(t,|x+y|). \tag{4.7c}$$

Then the heat kernel of L in $(0,b)$, with the same boundary conditions is (as in the previous sections)

$$k(t,x,y) = \sum_{n=0}^{\infty} k_n(t,x,y), \tag{4.8a}$$

where

$$k_0(t,x,y) = p_b^m(t,x,y) \tag{4.8b}$$

and, for $n \geq 1$,

$$k_n(t,x,y) = -\int_0^t \int_0^b p_b^m(s,x,z)q(z)k_{n-1}(t-s,z,y)dzds. \tag{4.8c}$$

In fact there is a Feynman–Kac type of expansion for $k(t,x,y)$ (see [H]).

In our regular notation, let $\mu_1, \mu_2, \ldots, \mu_n, \ldots$ be the eigenvalues of L on $(0,b)$ with corresponding (orthonormal) eigenfunctions $\phi_1, \phi_2, \ldots, \phi_n, \ldots$. The corresponding quantities for L_0 are $\nu_1, \nu_2, \ldots, \nu_n, \ldots$ and $\psi_1, \psi_2, \ldots, \psi_n, \ldots$. Of course

$$\nu_n = \left(\frac{\pi n}{b}\right)^{2m} \quad \text{and} \quad \psi_n(x) = \sqrt{\frac{2}{b}} \sin \frac{\pi n x}{b}, \qquad n = 1,2,\ldots \tag{4.9}$$

and

$$|\mu_n - \nu_n| \leq \|q\|_\infty. \tag{4.9'}$$

Then, in the Lemma of §1 we have $H(s) = L_0 + sq$ and so

$$\lambda_n(0) = \nu_n, \quad \lambda_n(1) = \mu_n, \quad \sup_{0 \leq s \leq 1} |\lambda_n''(s)| = O(n^{1-2m}),$$

by (1.7), (4.9), (4.9'), and the definition of $R_n(s)$ given after formula (1.7). Finally, from (1.6) we get

$$\lambda_n'(0) = \frac{2}{b}\int_0^b q(x) \sin^2 \frac{\pi n x}{b} dx = -\frac{1}{b}\int_0^b q(x) \cos \frac{2\pi n x}{b} dx,$$

where we have made our standard assumption, namely

$$\int_0^b q(x)dx = 0. \tag{4.10}$$

Thus, (1.8) and the Riemann–Lebesgue lemma imply

$$\mu_n - \nu_n = O(n^{-k \wedge (2m-1)}), \tag{4.11}$$

where k is such that $q \in C^k[0,b]$. If $m > 1$, (4.11) is quite satisfying, although we believe it can be improved.

Now we want to derive the trace formula for L.

For any $m \geq 2$, (4.8) and (4.7b) easily imply

(4.12)
$$k_0(t,x,y) = O(t^{-1/4}), \quad k_1(t,x,y) = O(t^{3/4}), \quad k_2(t,x,y) = O(t^{7/4}), \quad \text{as } t \downarrow 0,$$

uniformly in x, y. These are very crude estimates but we do not need any better for our analysis. As in §1, we have

$$\operatorname{tr}(e^{-tL} - e^{-tL_0}) = \sum_n (e^{-\mu_n t} - e^{-\nu_n t}) = \int_0^b k_1(t,x,x)dx + O(t^{7/4}).$$

Equation (4.8c) together with Fubini's theorem and the fact that $p_b^m(t,x,y)$ is the kernel of a semigroup give

$$\begin{aligned}\int_0^b k_1(t,x,x)dx &= -\int_0^b \int_0^t \int_0^b p_b^m(s,x,z)q(z)p_b^m(t-s,z,x)dzdsdx \\ &= -t\int_0^b p_b^m(t,z,z)q(z)dz.\end{aligned}$$

Therefore

$$\sum_n (e^{-\mu_n t} - e^{-\nu_n t}) = -t\int_0^b p_b^m(t,z,z)q(z)dz + O(t^{7/4}). \tag{4.13}$$

Now, from (4.7) we have for all $k > 0$, $0 \leq z \leq b$ (as in (1.11))

$$p_b^m(t,z,z) = \frac{C_1^m}{t^{1/2m}} - H_1^m(t,0,2z) - H_1^m(t,0,2b-2z) + o(t^k), \quad \text{as } t \downarrow 0. \tag{4.14}$$

LEMMA 4.2. *For any $f \in C_b(R^d)$ we have*

$$\lim_{t\downarrow 0} \int_{R^d} H_d^m(t,x,y)f(y)dy = f(x)$$

uniformly in x.

Proof. If f is in $L^2(R^d)$, then the statement is true in the L^2 sense, simply because H_d^m is the integral kernel of a semigroup. Furthermore, in the long remark following equation (4.4′) and in Lemma 4.1 right after this remark, we saw some nice estimates for H_d^m. These estimates together with the L^2 convergence immediately imply the statement of the lemma. □

Finally, if we combine (4.13) and (4.14) and then apply Lemma 4.2 and assume (4.10) (namely, that the average of q on $[0,b]$ is zero), we obtain

$$\sum_n (e^{-\mu_n t} - e^{-\nu_n t}) = -t\int_0^b p_b^m(t,z,z)q(z)dz + O(t^{5/4}),$$

for $q \in C^k[0,b]$ $k \geq 2$. Then, because of (4.11) we can divide by t and let $t \downarrow 0$ to conclude

$$\sum_n (\mu_n - \nu_n) = \frac{q(0)+q(b)}{4}, \tag{4.15}$$

by (4.14) and Lemma 4.2.

Remark. Formula (4.15) hints an interesting theoretical implication. It points out that the trace of the difference $[(-d^2/dx^2)^m + q] - (-d^2/dx^2)^m$ is independent of m and, therefore can be thought as the trace of the multiplication operator $(Tu)(x) = q(x)u(x)$. Of course q must be smooth and average-free. The boundary conditions play a role (in case of Neumann or Robin conditions there is a sign change in (4.15)). We believe that there are many things yet to be understood.

5. Higher-order Schrödinger operators II. The diophantine case. In this section we extend the results of §3. To be more precise, we obtain a trace formula and asymptotics of eigenvalues for

$$L = L_0 + q, \quad \text{where } L_0 = (-\Delta)^m$$

acts upon $L^2(D)$, with D being the rectangle $(0, b_1) \times (0, b_2)$ as in §3. It turns out that the asymptotics of the eigenvalues improve as m increases. Intuitively, this means that, as m gets bigger, L becomes a smaller perturbation of L_0.

First, let us decide what are our boundary conditions. The quantity

$$\begin{aligned} p_Q^m(t,x,y) = {} & H_2^m(t,x_1,x_2,y_1,y_2) - H_2^m(t,x_1,x_2,y_1,-y_2) \\ & - H_2^m(t,x_1,x_2,-y_1,y_2) + H_2^m(t,x_1,x_2,-y_1,-y_2), \end{aligned} \tag{5.1a}$$

where $H_2^m(t,x,y) = H_2^m(t,x_1,x_2,y_1,y_2)$ is defined in (4.4), is the heat kernel of L_0 acting upon the first quadrant $Q = \{(x_1,x_2) : x_1, x_2 > 0\}$. The boundary conditions are

$$\frac{\partial^{2j} p_Q^m}{\partial x_1^{2j}}(t,0,x_2,y_1,y_2) = 0 \quad \text{and} \quad \frac{\partial^{2j} p_Q^m}{\partial x_2^{2j}}(t,x_1,0,y_1,y_2) = 0, \qquad j = 0,1\ldots,m-1.$$

This follows from the fact that $e_d(x,y;\lambda)$ and therefore $H_d^m(t,x,y)$ are even and smooth in $|x-y|$, as explained in the beginning of §4. After this simple observation, it is easy to construct the heat kernel $p_D^m(t,x,y)$ of L_0 in $L^2(D)$ with "generalized Dirichlet" boundary conditions, namely vanishing normal derivatives of order $0, 2, \ldots, m-1$ on ∂D, from $H_2^m(t,x,y)$ by using the method of images (separation of variables fails if $m > 1$).

$$p_D^m(t,x,y) = \sum_{k \in Z^2} p_Q^m(t,x,y+2kb), \quad \text{where } b = (b_1,b_2), \tag{5.1b}$$

where the convergence is guaranteed by the long remark after equation (4.4′).

Assuming (3.11), which says that the average of q along each segment parallel to the sides of the rectangle D is 0, we arrive (by the standard heat kernel approach) at a formula very similar to (3.12) thanks to Lemma 4.2, namely

$$\lim_{t \downarrow 0} \frac{1}{t} \sum_n (e^{-t\mu_n} - e^{-t\nu_n}) = -\frac{1}{16}[q(0,0) + q(b_1,0) + q(0,b_2) + q(b_1,b_2)]. \tag{5.2}$$

The most interesting result of this section is that in certain cases we are able to justify the interchange of limit and summation in (5.2). It is a direct consequence of the following proposition.

PROPOSITION. *Let $\omega = b_1^2/b_2^2$ be irrational and such that, for some $\alpha \geq 2$, $c > 0$, the inequality $|\omega - p/q| > cq^{-\alpha}$, where p and q are positive integers, is true for all but finitely many q (in this case ω is called diophantine; in fact, the set of nondiophantine numbers has Lebesgue measure zero). Then there is a constant $A > 0$ such that, for all n sufficiently large, we have $\nu_{n+1} - \nu_n > An^{m-\alpha}$, where ν_n is the nth ("generalized Dirichlet") eigenvalue of $L_0 = (-\Delta)^m$ on $D = (0, b_1) \times (0, b_2)$, namely*

$$\nu_n = \left(\frac{\pi^2 n_1^2}{b_1^2} + \frac{\pi^2 n_2^2}{b_2^2} \right)^m, \qquad n_1, n_2 = 1, 2, 3, \ldots$$

and the labeling is done in the unique way that makes ν_n increasing with n.

Proof. Set

$$\tau_n = \frac{\pi^2 n_1^2}{b_1^2} + \frac{\pi^2 n_2^2}{b_2^2} \quad \text{and} \quad \tau_{n+1} = \frac{\pi^2 k_1^2}{b_1^2} + \frac{\pi^2 k_2^2}{b_2^2}.$$

If n is such that $n_2 = k_2$, then $k_1 = n_1 + 1$ and the theorem is trivially true. Hence, let us assume that $n_2 \neq k_2$ so that

$$\tau_{n+1} - \tau_n = \frac{\pi^2 |n_2^2 - k_2^2|}{b_1^2} \left| \omega - \frac{k_1^2 - n_2^2}{n_2^2 - k_2^2} \right|.$$

Then, our assumption for ω together with (3.2b) imply (since, for fixed q, the equation $n_2^2 - k_2^2 = q$ has finitely many integer solutions)

$$\tau_{n+1} - \tau_n > \frac{\pi^2}{b_1^2} \frac{c}{\max(n_2^2, k_2^2)^{\alpha - 1}} > \frac{C}{n^{\alpha - 1}}, \tag{5.3}$$

where C is a constant. Finally, since

$$\nu_{n+1} - \nu_n = \tau_{n+1}^m - \tau_n^m = (\tau_{n+1} - \tau_n)(\tau_{n+1}^{m-1} + \cdots + \tau_n^{m-1}),$$

the proof is completed by using (5.3) and (3.2b). □

Remark. If ω is algebraic then any $\alpha > 2$ works. This is the famous Thue–Siegel–Roth Theorem. In fact, the set of all reals for which this is not true has measure zero (see [L]).

COROLLARY. *If μ_n is the nth eigenvalue of $L = L_0 + q$, with $q \in C^r(\bar{D})$ satisfying (3.11), and everything else is as in the previous proposition, then there is a constant $c > 0$ such that*

$$|\mu_n - \nu_n| \leq \frac{c}{n^\beta}, \quad \beta = \min(r/2, m - \alpha), \quad \textit{for all } n \textit{ sufficiently large.}$$

Proof. The notation is the same as in the Lemma of §1; namely, we set

$$H(s) = L_0 + sq \quad \text{with spectrum } \{\lambda_n(s)\}_{n=1}^\infty$$

and

$$R_n(s) = [I - P_n(s)][H(s) - \lambda_n(s)I]^{-1},$$

where $P_n(s)$ is the orthogonal projection on the (one-dimensional) eigenspace which corresponds to $\lambda_n(s)$. Thus, because of the previous proposition, if n is large enough then

$$\|R_n(0)\| \leq \frac{c}{n^{m-\alpha}},$$

with c being a constant. The normalized eigenfunction $\psi_n(x)$ of L_0 is given in (3.2a). Then (1.7) implies that

$$|\lambda_n''(0)| \leq \frac{c}{n^{m-\alpha}} \quad \text{for all } n \text{ sufficiently large.}$$

Also, (1.6) and (3.11) imply

$$\begin{aligned}\lambda_n'(0) &= \int_D q(x)\psi_n(x)^2 dx \\ &= \frac{1}{\sqrt{b_1 b_2}} \int_D q(x) \cos\frac{2n_1 x_1}{b_1} \cos\frac{2n_2 x_2}{b_2} dx = O(\max(n_1, n_2)^{-r}) = O(n^{-r/2}),\end{aligned}$$

as $n \to \infty$ and, therefore, the corollary follows from (1.8). □

Using this corollary we can conclude that, if $r > 2$ and $m - \alpha > 1$ (in particular, if $(b_1/b_2)^2$ is an algebraic irrational and $m \geq 4$), we can bring the limit in (5.2) inside the summation and deduce

$$\sum_n (\mu_n - \nu_n) = \frac{1}{16}[q(0,0) + q(b_1,0) + q(0,b_2) + q(b_1,b_2)]. \tag{5.4}$$

The remark at the end of the previous section applies here as well.

6. Final thoughts (epilogue). One of our main tasks in the previous sections was to find short-time asymptotics for the trace

$$\operatorname{tr}\left(e^{-tL} - e^{-tL_0}\right),$$

where L and L_0 were certain self-adjoint differential operators, L being a perturbation of L, namely $L = L_0 + q$. These asymptotics involved the values of q on the boundary of the domain on which these operators were acting. On the other hand, this trace can be written as a sum of the differences of the eigenvalues $\{e^{t\mu_n}\}$ and $\{e^{-t\nu_n}\}$ of e^{-tL} and e^{-tL_0}, respectively. Then we could take the limits as $t \downarrow 0$. In some cases we were able to interchange limit and trace (i.e., summation) and thus arrive to a formula for

$$\sum_n (\mu_n - \nu_n),$$

which, in some sense, is the trace of $L - L_0$. In the other cases the problem remains open, namely, can we (at least in some "weak" sense) interchange these limiting procedures?

The approach we followed can be applied to more general situations. Here is a final example.

Let L_0 be the operator $-\Delta$ acting upon the rectangle $D = (0, b_1) \times (0, b_2)$ with Neumann boundary conditions (i.e., the normal derivative on ∂D is 0) and L be again the Laplacian operator acting upon D, but this time the boundary condition is

$$\frac{\partial u}{\partial n} + \rho u = 0 \quad \text{on } \partial D,$$

where n is the outward unit normal vector on ∂D and ρ is a smooth function on ∂D *with zero average on each segment of the perimeter of* D. The heat kernel of L_0 is

$$\bar{p}_D(t, x, y) = \bar{p}_{b_1}(t, x_1, y_1)\bar{p}_{b_2}(t, x_2, y_2),$$

where

$$\bar{p}_b(t,x,y) = \sum_{k\in Z} \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{|x-y+2kb|^2}{4t}\right) + \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{|x+y+2kb|^2}{4t}\right)$$

(p_b is the Neuman heat kernel of $-d^2/dx^2$ acting on $(0,b)$). The heat kernel $k(t,x,y)$ of L is given in [P]. Using the same analysis as in the previous cases we arrive at the formula

$$\lim_{t\downarrow 0} \frac{1}{\sqrt{t}} \sum_n (e^{-t\mu_n} - e^{-t\nu_n})$$

$$= \lim_{t\downarrow 0} \sqrt{t} \int_{\partial D} p_D^m(t,z,z)\rho(z)dz = -\frac{1}{4\sqrt{\pi}}[\rho(0,0) + \rho(b_1,0) + \rho(0,b_2) + \rho(b_1,b_2)].$$

Here, of course, there is no chance that we can interchange limit and summation unless $\rho(0,0) + \rho(b_1,0) + \rho(0,b_2) + \rho(b_1,b_2) = 0$.

Acknowledgments. We thank Stephanos Venakides for initiating the whole inquiry and for many helpful discussions. Also we thank Peter Kuchment for asking instructive questions, Walter Craig for showing to us A. Douady's insightful lecture on the trace formula, and finally Christopher Jones for helpful suggestions that made the paper more readable.

REFERENCES

[B] V. Barcilon, *Explicit solution of the inverse problem for a vibrating string*, J. Math. Anal. Appl., 93 (1983), pp. 222–234.

[B-O] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[C] W. Craig, *The trace formula for Schrödinger operators on the line*, Comm. Math. Phys., 126 (1989), pp. 379–407.

[C-H] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 1., Interscience, New York, 1953.

[D-H-V] A. Douady, J. H. Hubbard, and J. L. Verdier, *Equation de Hill periodique*, Seminaire de Geometrie Analytique de l'Ecole Normale Superiure 1976–77, seminar of A. Douady, Lecture Notes.

[F] L. Friedlander, *On the spectrum of the periodic problem for the Schrödinger operator*, Comm. Partial Differential Equations, 15 (1990), pp. 1631–1647.

[F-K-T] J. Feldman, H. Knorrer, and E. Trubowitz, *The perturbative stable spectrum of a periodic Schrödinger operator*, Invent. Math., 100 (1990), pp. 259–300.

[G-L] I. M. Gelfand and B. M. Levitan, *On a simple identity for the eigenvalues of a differential equation*, Dokl. Akad. Nauk SSSR, 88 (1953), pp. 593–596.

[H] K. J. Hochberg, *A signed measure on path space related to wiener measure*, Ann. Probab., 6 (1978), pp. 433–458.

[H.H] H. Hochstadt, *Estimates on the stability intervals for Hill's equation*, Proc. Amer. Math. Soc., 14 (1963), pp. 930–932.

[J] F. John, *Partial Differential Equations*, fourth ed., Appl. Math. Sci. 1, Springer-Verlag, New York.

[L] S. Lang, *Introduction to Diophantine Approximations*, Addison-Wesley, New York, 1966.

[L-G] B. M. Levitan and M. G. Gasymov, *Determination of a differential equation by two of its spectra*, Russian Math. Surveys, 19 (1964), pp. 1–63.

[L-S] B. M. Levitan and I. S. Sargsjan, *Introduction to Spectral Theory: Self Adjoint Ordinary Differential Operators*, Am. Math. Soc. Transl. 39, American Mathematical Society, Providence, RI.

[P] V. G. Papanicolaou, *The probabilistic solution of the third boundary value problem for second order elliptic equations*, Probab. Theory Related Fields, 87 (1990), pp. 27–77.

[R-S] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* IV, *Analysis of Operators*, Academic Press, New York, 1978.

[S] B. Simon, *Schrödinger Semigroups*, Bull. Amer. Math. Soc., 7 (1982), pp. 447–526.

SIAM J. MATH. ANAL.
Vol. 26, No. 1, pp. 238–262, January 1995

012

NECESSARY AND SUFFICIENT CONDITIONS FOR MEAN CONVERGENCE OF LAGRANGE INTERPOLATION FOR FREUD WEIGHTS *

D. S. LUBINSKY† AND D. M. MATJILA‡

Abstract. Let $W_\beta(x) := \exp(-\frac{1}{2}|x|^\beta), x \in \mathbb{R}, \beta > 1$. Given $f : \mathbb{R} \to \mathbb{R}$, let $L_n[f](x)$ denote the Lagrange interpolation polynomial to f at the zeros of the orthonormal polynomial of degree n for the weight W_β^2. Let $1 < p < \infty, \Delta \in \mathbb{R}, \alpha > 0$, and $\hat{\alpha} := \min\{1, \alpha\}$. Moreover, let

$$\tau = \tau(p) := 1/p - \hat{\alpha} + \begin{cases} 0, & p \le 4, \\ (\beta/6)(1 - 4/p), & p > 4. \end{cases}$$

It is shown that for

$$\lim_{n\to\infty} \| (f(x) - L_n[f](x)) W(x)(1 + |x|)^{-\Delta} \|_{L_p(\mathbb{R})} = 0, \tag{1}$$

to hold for every continuous function $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$\lim_{|x|\to\infty} |f(x)| W(x)(1 + |x|)^\alpha = 0, \tag{2}$$

it is necessary and sufficient that

$$\Delta > \tau \quad \text{if } 1 < p \le 4;$$
$$\Delta > \tau \quad \text{if } p > 4 \quad \text{and} \quad \alpha = 1;$$
$$\Delta \ge \tau \quad \text{if } p > 4 \quad \text{and} \quad \alpha \ne 1.$$

Moreover, it is shown that (1) holds for every $1 < p < \infty$ and every continuous function f satisfying (2) if and only if $\Delta \ge -\hat{\alpha} + \max\{1, \beta/6\}$. These are special cases of results for more general Freud weights.

Key words. Freud weights, Lagrange interpolation, mean convergence, L_p norms

AMS subject classifications. primary 42C05, 42C15; secondary 65D05

1. Introduction and results.

The convergence of Lagrange interpolation at zeros of orthogonal polynomials is a classical and widely studied subject. Let us recall the setting. If $d\alpha$ is a finite positive Borel measure on a finite or infinite interval (a, b), with support containing infinitely many points, and with all power moments

$$\int_a^b x^j \, d\alpha(x), j = 0, 1, 2, \ldots,$$

* Received by the editors September 14, 1992; accepted for publication (in revised form) June 9, 1993.

† Department of Mathematics, University of the Witwatersrand, P.O. Wits 2050, Republic of South Africa.

‡ Department of Mathematics, University of the North, P.O. Box X1106, Sovenga 0727, Republic of South Africa.

finite, then we may define orthonormal polynomials $p_n(d\alpha, x)$ (or simply $p_n(x)), n = 0, 1, 2, \ldots$, which satisfy

$$\int_a^b p_n(d\alpha, x)p_m(d\alpha, x)\, d\alpha(x) = \delta_{mn}, \qquad m, n = 0, 1, 2, \ldots.$$

We denote the zeros of $p_n(d\alpha, x)$ by

$$a < x_{nn} < x_{n-1,n} < \cdots < x_{2n} < x_{1n} < b.$$

For functions $f : (a, b) \to \mathbb{R}$, we denote the *Lagrange interpolation polynomial* of degree $\leq n - 1$ to f at the zeros of $p_n(d\alpha, x)$ by $L_n[f](x)$, so that

$$L_n[f](x_{jn}) = f(x_{jn}), \qquad 1 \leq j \leq n. \tag{1.1}$$

The classical Erdös–Turan theorem asserts that if (a, b) is a finite interval, then for each measurable function $f : (a, b) \to \mathbb{R}$ for which

$$\int_a^b f^2\, d\alpha < \infty,$$

we have L_2 convergence:

$$\lim_{n\to\infty} \int_a^b (f - L_n[f])^2\, d\alpha = 0.$$

The extension of the Erdös–Turan theorem to infinite intervals, is due primarily to Shohat; see [3, Chap. III]. The more difficult extension of the Erdös–Turan theorem on a finite interval (a, b) to the $L_p, p \neq 2$ case has attracted many authors and inspired research into fundamental properties of orthogonal polynomials. The results obtained are inherently more special, and necessarily require more knowledge of the orthonormal polynomials than in the L_2 case. We cannot hope to survey this topic here; the reader may find results and references in [12], [14], and [15].

In this paper, we concentrate on weights on the whole real line. We shall deal with "Freud weights," that is, weights of the form

$$d\alpha(x) = W^2(x)dx,$$

where $W : \mathbb{R} \to \mathbb{R}$ is even and of sufficiently smooth and regular decay at infinity. (Note that our weight is of the form W^2, *not* W: This simplifies formulation of results.) Here we shall denote $p_n(d\alpha, x)$ by $p_n(W^2, x), n = 0, 1, 2, \ldots$. Our results apply in particular to

$$W(x) = W_\beta(x) := \exp(-\tfrac{1}{2}|x|^\beta), \qquad x \in \mathbb{R}, \quad \beta > 1. \tag{1.2}$$

Many of our methods are taken from a fundamental paper of Nevai [13], which dealt with the Hermite weight, the case $\beta = 2$ in (1.2). Nevai provided fairly close necessary and sufficient conditions for L_p convergence of the Lagrange interpolation polynomials, in terms of the decay of the interpolated function. Nevai's methods pointed the way of subsequent research: One of Nevai's students, Bonan, obtained "sharp" necessary and sufficient conditions for the generalized Hermite weights $|x|^\beta \exp(-x^2)$,

in an unpublished Ph.D. thesis [1]. His decay condition on the interpolated function was, however, a little more restrictive than Nevai's.

Mean convergence of Lagrange interpolation is dependent on suitable estimates for associated orthonormal polynomials, and so it is not surprising that Nevai and Bonan at the time concentrated on the Hermite weight, for which the relevant estimates and Plancherel type asymptotics were available. By assuming bounds on orthonormal polynomials that were known to be true for the weights W_β^2, β a positive even integer, Knopfmacher and Lubinsky treated fairly general Freud weights $W^2 = e^{-2Q}$ [4]. For the L_p norm with $1 < p < 2$, the results there sharpened and generalised those in [1], and functions with integrable singularities at finitely many points were also discussed, which extended L_2 work in [8].

The possibility of treating weights such as $W_\beta^2, \beta > 1$ is provided by work of Levin and Lubinsky [5], where the correct bounds were obtained for orthonormal polynomials over $\mathbb{R}$ for a large class of weights. We also use the L_p norms of orthonormal polynomials, and these were estimated above and below in [7], using the results of [5].

The following is an important special case of our results.

THEOREM 1.1. *Let $\beta > 1, 1 < p < \infty, \Delta \in \mathbb{R}, \alpha > 0$, and*

$$\hat{\alpha} := \min\{1, \alpha\}. \tag{1.3}$$

Let

$$\tau = \tau(p) := \begin{cases} \frac{1}{p} - \hat{\alpha}, & p \le 4, \\ \frac{1}{p} - \hat{\alpha} + \frac{\beta}{6}\left(1 - \frac{4}{p}\right), & p > 4. \end{cases} \tag{1.4}$$

Denote the Lagrange interpolation polynomials at the zeros of $p_n(W_\beta^2, \cdot)$ by $L_n[\cdot], n \ge 1$, where W_β is given by (1.2). *Then for*

$$\lim_{n\to\infty} \|(f(x) - L_n[f](x))W_\beta(x)(1 + |x|)^{-\Delta}\|_{L_p(\mathbb{R})} = 0, \tag{1.5}$$

to hold for every continuous function $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$\lim_{|x|\to\infty} |f(x)|W_\beta(x)(1 + |x|)^\alpha = 0, \tag{1.6}$$

it is necessary and sufficient that

$$\begin{aligned} &\Delta > \tau \quad \text{if} \quad 1 < p \le 4; \\ &\Delta > \tau \quad \text{if} \quad p > 4 \quad \text{and} \quad \alpha = 1; \\ &\Delta \ge \tau \quad \text{if} \quad p > 4 \quad \text{and} \quad \alpha \ne 1. \end{aligned}$$

Remarks. (a) In the work of Bonan [1], the Hermite weight ($\beta = 2$) was treated, and it was assumed that $\alpha = 2$. In this special case, our sufficient condition is identical to Bonan's, and our necessary condition is also identical if we take the $v(x) = W_\beta(x)(1 + |x|)^{-\Delta}$ in Bonan's work.

(b) The sufficient condition also guarantees the convergence (1.5) if f is not necessarily continuous but is bounded and Riemann integrable on each finite interval and satisfies (1.6). However, we shall not prove this, as it is away from the focus of this paper, which is sharp conditions relating α and Δ. The reader may refer to Bonan's thesis [1] or Knopfmacher and Lubinsky [4].

(c) Results such as the previous ones are useful in investigating convergence of product integration rules (cf. [4]).

(d) When we consider convergence simultaneously for all $1 < p < \infty$, the above result simplifies substantially.

COROLLARY 1.2. *Let $\beta > 1, \Delta \in \mathbb{R}, \alpha > 0$, and $\hat{\alpha} := \min\{1, \alpha\}$. For the convergence* (1.5) *to hold for every $1 < p < \infty$ and every continuous function $f : \mathbb{R} \to \mathbb{R}$ satisfying* (1.6), *it is necessary and sufficient that*

$$\Delta \geq -\hat{\alpha} + \max\left\{1, \frac{\beta}{6}\right\}. \tag{1.7}$$

To formulate our result for more general weights, we need the Mhaskar–Rahmanov–Saff number a_u [9], [10]. Let $W := e^{-Q}$, where $Q : \mathbb{R} \to \mathbb{R}$ is even, continuous, and $xQ'(x)$ is positive and increasing in $(0, \infty)$, with limits 0 and ∞ at 0 and ∞, respectively. For $u > 0$, the uth Mhaskar–Rahmanov–Saff number a_u is the positive root of the equation

$$u = \frac{2}{\pi} \int_0^1 a_u t Q'(a_u t)\, dt/\sqrt{1 - t^2}. \tag{1.8}$$

Under the conditions on Q below, which guarantee that $Q(s)$ and $Q'(s)$ increase strictly in $(0, \infty)$, a_u is uniquely defined and increases with u. It grows roughly like $Q^{[-1]}(u)$, where $Q^{[-1]}$ denotes the inverse of Q on $(0, \infty)$. Its significance lies partly in the identity

$$\|PW\|_{L_\infty(\mathbb{R})} = \|PW\|_{L_\infty[-a_n, a_n]}, \tag{1.9}$$

which holds for polynomials P of degree $\leq n$ [10]. For $W = W_\beta$, one sees that for $u > 0$,

$$a_u = Cu^{1/\beta},$$

where C depends only on β.

Theorem 1.1 is a special case of the following theorem.

THEOREM 1.3. *Let $W := e^{-Q}$, where $Q : \mathbb{R} \to \mathbb{R}$ is even and continuous in $\mathbb{R}$, Q'' is continuous in $(0, \infty)$, and $Q' > 0$ in $(0, \infty)$, while for some $A, B > 1$,*

$$A \leq \frac{d}{dx}(xQ'(x))/Q'(x) \leq B, \qquad x \in (0, \infty). \tag{1.10}$$

Let $1 < p < \infty, \Delta \in \mathbb{R}, \alpha > 0$, and $\hat{\alpha}$ be given by (1.3). *Denote the Lagrange interpolation polynomials at the zeros of $p_n(W^2, \cdot)$ by $L_n[\cdot], n \geq 1$. Then for*

$$\lim_{n \to \infty} \|(f(x) - L_n[f](x))W(x)(1 + |x|)^{-\Delta}\|_{L_p(\mathbb{R})} = 0, \tag{1.11}$$

to hold for every continuous function $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$\lim_{|x| \to \infty} |f(x)|W(x)(1 + |x|)^{\alpha} = 0, \tag{1.12}$$

if $p \leq 4$, it is necessary and sufficient that

$$\Delta > -\hat{\alpha} + \frac{1}{p}; \tag{1.13}$$

and if $p > 4$ *and* $\alpha \neq 1$, *it is necessary and sufficient that*

$$a_n^{1/p-(\hat{\alpha}+\Delta)} n^{(1/6)(1-4/p)} = O(1), \quad n \to \infty; \tag{1.14}$$

and if $p > 4$ *and* $\alpha = 1$, *it is necessary and sufficient that*

$$a_n^{1/p-(\hat{\alpha}+\Delta)} n^{(1/6)(1-4/p)} = O\left(\frac{1}{\log n}\right), \quad n \to \infty. \tag{1.15}$$

The paper is organised as follows: In §2, we present some technical lemmas, such as estimates on orthonormal polynomials and their zeros, and we use these to prove three quadrature sum estimates, including a very technical one (Lemma 2.7), which is essentially a bound on part of the Lebesgue function of Lagrange interpolation.

In §3, we prove the results. The proof of the sufficiency part of Theorem 1.3 involves expressing the given function f, for a given n, as a sum of functions, one vanishing outside $[-a_n/4, a_n/4]$ and another vanishing inside $(-a_n/4, a_n/4)$ and approximating f by a suitable polynomial. This is achieved in Lemmas 3.1–3.4. The proof of the necessary conditions involves the uniform boundedness principle applied to carefully chosen spaces, and interpolation of carefully chosen functions.

2. Technical lemmas. We first need more notation. Throughout, $C, C_1, C_2, \ldots$, denote positive constants independent of n, x, and polynomials of degree n. We shall write $C \neq C(\lambda)$ to indicate that C does not depend on a parameter λ. The same symbol does not necessarily represent the same constant in different occurrences. We use $\sim$ in the following sense: If $\{b_n\}_{n=0}^{\infty}$ and $\{c_n\}_{n=0}^{\infty}$ are sequences of nonzero real numbers, we write

$$b_n \sim c_n,$$

if there exist $C_1, C_2 > 0$ independent of n, such that

$$C_1 \leq b_n/c_n \leq C_2, \qquad n \geq 1.$$

Similar notation is used for sequences of functions. We frequently use $f^n(x)$ to denote $(f(x))^n$, even for $n = -1$. $\mathcal{P}_n$ denotes the polynomials of degree at most n.

If $W^2 : \mathbb{R} \to [0, \infty)$ is a weight, its nth *Christoffel function* is

$$\lambda_n(x) := \lambda_n(W^2, x) := \inf_{P \in \mathcal{P}_{n-1}} \int_{-\infty}^{\infty} (PW)^2(t)\, dt / P^2(x). \tag{2.1}$$

It is well known [14] that

$$\lambda_n(x) = 1 \Big/ \sum_{j=0}^{n-1} p_j^2(x).$$

We also define the *Christoffel numbers*

$$\lambda_{jn} := \lambda_n(x_{jn}), \qquad 1 \leq j \leq n.$$

The leading coefficient of $p_n(x)$ is denoted by γ_n, so that

$$p_n(x) = \gamma_n x^n + \cdots.$$

The Lagrange interpolation polynomial $L_n[f]$ admits the representation

$$L_n[f](x) = \sum_{j=1}^{n} f(x_{jn})\ell_{jn}(x),$$

where the *fundamental polynomials* ℓ_{jn} in turn admit the representation [12, p. 6] or [3, p. 23]

$$\ell_{jn}(x) = \lambda_{jn}\frac{\gamma_{n-1}}{\gamma_n}p_{n-1}(x_{jn})\frac{p_n(x)}{x - x_{jn}}. \tag{2.2}$$

Mean convergence of Lagrange interpolation is closely connected to bounds on orthogonal polynomials and related estimates; accordingly we recall some results from [5]. Throughout this section, we assume that W is as in Theorem 1.3.

THEOREM 2.1. (a) *For $n \geq 1$ and $|x| \leq a_n$,*

$$\lambda_n(W^2, x) \sim \frac{a_n}{n}W^2(x)\max\left\{n^{-2/3}, 1 - \frac{|x|}{a_n}\right\}^{-1/2}. \tag{2.3}$$

Moreover, we can replace $\sim$ by $\geq C \times$ for all $x \in \mathbb{R}$.

(b) *For $n \geq 1$,*

$$|x_{1n}/a_n - 1| \leq Cn^{-2/3}, \tag{2.4}$$

and uniformly for $n \geq 3$ and $2 \leq j \leq n-1$,

$$x_{j-1,n} - x_{j+1,n} \sim \frac{a_n}{n}\max\{n^{-2/3}, 1 - |x_{jn}|/a_n\}^{-1/2}. \tag{2.5}$$

(c) *For $n \geq 1$,*

$$\sup_{x\in\mathbb{R}}|p_n(x)|W(x)|1 - |x|/a_n|^{1/4} \sim a_n^{-1/2}, \tag{2.6}$$

and

$$\sup_{x\in\mathbb{R}}|p_n(x)|W(x) \sim n^{1/6}a_n^{-1/2}. \tag{2.7}$$

(d) *Let $0 < p \leq \infty$. There exists $C > 0$ such that for $n \geq 1$ and $P \in P_n$,*

$$\|PW\|_{L_p(\mathbb{R})} \leq C\|PW\|_{L_p[-a_n,a_n]}. \tag{2.8}$$

(e) *For $n \geq 1$,*

$$\gamma_{n-1}/\gamma_n \sim a_n. \tag{2.9}$$

(f) *Uniformly for $n \geq 2$ and $1 \leq j \leq n-1$,*

$$\max\{n^{-2/3}, 1 - |x_{jn}|/a_n\} \sim \max\{n^{-2/3}, 1 - |x_{j+1,n}|/a_n\}.$$

Proof. (a)–(d) are, respectively, Theorem 1.1, Corollaries 1.2 and 1.4, and Theorem 1.8 in [5]. (e) is Theorem 12.3(b) in [5]. (f) is (1.10) in [5, p. 521]. □

We recall from [7] the next theorem.

THEOREM 2.2. (a) *Given* $0 < p < \infty$, *we have for* $n \geq 1$,

$$\|p_n W\|_{L_p(\mathbb{R})} \sim a_n^{1/p-1/2} \times \begin{cases} 1, & p < 4, \\ (\log n)^{1/4}, & p = 4, \\ n^{(1/6)(1-4/p)}, & p > 4. \end{cases} \tag{2.10}$$

(b) *Uniformly for* $n \geq 1, 1 \leq j \leq n$, *and* $x \in \mathbb{R}$,

$$|\ell_{jn}(x)| \sim \frac{a_n^{3/2}}{n} W(x_{jn}) \left(\max \left\{ n^{-2/3}, 1 - \frac{|x_{jn}|}{a_n} \right\} \right)^{-1/4} \left| \frac{p_n(x)}{x - x_{jn}} \right|. \tag{2.11}$$

(c) *Uniformly for* $n \geq 1, 1 \leq j \leq n$, *and* $x \in \mathbb{R}$,

$$|\ell_{jn}(x)| W^{-1}(x_{jn}) W(x) \leq C. \tag{2.12}$$

Proof. (a), (b), and (c) are, respectively, Theorem 1 and Lemma 2.6(a), (b) in [7]. □

LEMMA 2.3. (a) *Given* $0 < a < b$, *we have uniformly for* $n \geq 1$ *and* $x \in [a, b]$,

$$Q(a_n x) \sim a_n x Q'(a_n x) \sim n. \tag{2.13}$$

(b) *If* A, B *are as in* (1.10), *then*

$$u^{1/B} \leq a_u / a_1 \leq u^{1/A}, \qquad u \in [1, \infty). \tag{2.14}$$

(c) *Given* $\lambda > 1$, *we have for* $v \in (0, \infty)$, *and* $u \in [v/\lambda, \lambda v]$,

$$|a_u / a_v - 1| \sim |u/v - 1|. \tag{2.15}$$

Proof. (a) This is Lemma 5.1(c) in [5].
(b) This is Lemma 5.2(b) in [5].
(c) This is Lemma 5.2(c) in [5]. □

There is an old result of Shohat [3, Chap. III] that establishes the equivalence of convergence of Gauss quadratures and L_2 convergence of Lagrange interpolation. So it is not surprising that quadrature sums play a role here. In the remainder of this section, we prove three quadrature sum estimates. The first is a quadrature sum similar to that in Theorem 6 in [4, p. 85]. In the sequel, we set

$$x_{0n} := x_{1n}(1 + n^{-2/3}); \qquad x_{n+1,n} := x_{nn}(1 + n^{-2/3}). \tag{2.16}$$

LEMMA 2.4. *Let* $\nu \in \mathbb{R}$. *Then uniformly for* $n \geq 1$,

$$\sum_{j=1}^{n} \lambda_{jn} W^{-2}(x_{jn})(1 + |x_{jn}|)^{\nu} \sim \begin{cases} 1, & \nu < -1, \\ \log n, & \nu = -1, \\ a_n^{1+\nu}, & \nu > -1. \end{cases} \tag{2.17}$$

Proof. We first show that uniformly for $n \geq 1$ and $1 \leq j \leq n$,

$$1 + |x_{jn}| \sim 1 + |t|, \qquad t \in [x_{j+1,n}, x_{j-1,n}]. \tag{2.18}$$

To this end, note that (2.5) holds even for $j = 1$ and n, with the definition (2.16) of x_{0n} and $x_{n+1,n}$. First, if $|x_{j\pm1,n}| \le a_n/2$, then (2.5) shows that

$$x_{j-1,n} - x_{j+1,n} \sim a_n/n,$$

so that for t in the range (2.18),

$$\left|\frac{1+|t|}{1+|x_{jn}|} - 1\right| \le \frac{|t-x_{jn}|}{1+|x_{jn}|} \le x_{j-1,n} - x_{j+1,n} \le C_1 a_n/n \to 0, \qquad n \to \infty.$$

(See (2.14) and recall that $A > 1$.) On the other hand, if $|x_{j\pm1,n}| > a_n/2$, then (2.5) shows that

$$\begin{aligned}\left|\frac{1+|t|}{1+|x_{jn}|} - 1\right| &\le \frac{|t-x_{jn}|}{1+|x_{jn}|} \\ &\le C_2(x_{j-1,n} - x_{j+1,n})/a_n \le C_3\frac{1}{n}(\max\{n^{-2/3}, 1-|x_{jn}|/a_n\})^{-1/2} \\ &\le C_4 n^{-2/3} \to 0 \quad \text{as } n \to \infty.\end{aligned}$$

So (2.18) holds in all cases. Next, (2.3) and (2.5) imply that

$$\lambda_{jn} W^{-2}(x_{jn}) \sim x_{j-1,n} - x_{j+1,n}$$

for $2 \le j \le n-1$. By (2.16) this holds for $j = 1, n$ also. So for $1 \le j \le n$,

$$\begin{aligned}\lambda_{jn} W^{-2}(x_{jn})(1+|x_{jn}|)^\nu &\sim (x_{j-1,n} - x_{j+1,n})(1+|x_{jn}|)^\nu \\ &\sim \int_{x_{j+1,n}}^{x_{j-1,n}} (1+|t|)^\nu \, dt.\end{aligned}$$

Summing for $j = 1$ to n, and using

$$x_{0n} \sim -x_{n+1,n} \sim a_n,$$

we obtain (2.17). □

In several places, we shall need to replace $(1+t^2)^\nu$ by an equivalent polynomial on a suitable interval. This is achieved in the following lemma.

LEMMA 2.5. *Let $\nu \in \mathbb{R}$. There exists $C > 0$ such that for $\lambda \ge 2$, there exist polynomials $\hat{P}_\lambda$ of degree $\le C\lambda \log \lambda$ such that*

$$\hat{P}_\lambda(t) \sim (1+t^2)^\nu, \tag{2.19}$$

uniformly for $t \in [-\lambda, \lambda]$ and $\lambda \ge 2$.

Proof. Let $f(t) := \log(1+(\lambda t)^2), t \in [-1,1]$. Then

$$|f'(t)| = \frac{2\lambda^2|t|}{1+(\lambda t)^2} \le \lambda, \qquad t \in [-1,1].$$

By Jackson's theorem, there exists a polynomial R_λ of degree at most λ such that

$$\|f - R_\lambda\|_{L_\infty[-1,1]} \le C_1 \|f'\|_{L_\infty[-1,1]}/\lambda \le C_1. \tag{2.20}$$

Here C_1 is an absolute constant. Next, let

$$S_n(u) := \sum_{j=0}^{n} u^j/j!$$

be the $(n+1)$st partial sum of the Maclaurin series of e^u. It is well known and easy to see that

$$S_n(u) \sim e^u, \qquad |u| \le n/9.$$

Next, from (2.20),

$$\|R_\lambda\|_{L_\infty[-1,1]} \le C_1 + \|f\|_{L_\infty[-1,1]} \le C_1 + \log(1+\lambda^2) \le C_2 \log \lambda.$$

Let $[x]$ denote the greatest integer $\le x$. Then, if we set

$$\hat{P}_\lambda(t) := S_{10[|\nu|C_2 \log \lambda]}(\nu R_\lambda(t/\lambda)),$$

we have for $|\nu|C_2 \log \lambda \ge 1$ and $t \in [-\lambda, \lambda]$ that

$$\hat{P}_\lambda(t) \sim \exp(\nu R_\lambda(t/\lambda)) \sim \exp(\nu f(t/\lambda)) = (1+t^2)^\nu.$$

For small λ, we can set $\hat{P}_\lambda := 1$; Finally, $\hat{P}_\lambda$ has degree $\le 10|\nu|C_2 \lambda \log \lambda$. □

We can now prove our main quadrature sum estimate.

LEMMA 2.6. *Fix $\sigma \in (0,1)$ and an integer $L \ge 3$, and let $\nu \in \mathbb{R}$. Then for $P \in \mathcal{P}_{Ln}$, and $n \ge 1$,*

$$\sum_{|x_{jn}| \le \sigma a_n} \lambda_{jn} |P(x_{jn})| W^{-1}(x_{jn})(1+x_{jn}^2)^\nu \le C_1 \int_{-a_{Ln}}^{a_{Ln}} |PW|(t)(1+t^2)^\nu \, dt. \tag{2.21}$$

Here $C_1 \ne C_1(P,n)$.

Proof. Our method follows that of [6]. Now by definition of the Christoffel function, and by Theorem 2.1(a), we have for $P \in \mathcal{P}_n$ and $x \in \mathbb{R}$ that

$$(PW)^2(x) \le C_2 \frac{n}{a_n} \int_{-\infty}^{\infty} (PW)^2(t) \, dt.$$

It follows that

$$\|PW\|_{L_\infty(\mathbb{R})}^2 \le C_2 \frac{n}{a_n} \|PW\|_{L_\infty(\mathbb{R})} \int_{-\infty}^{\infty} |PW|(t) \, dt,$$

and so, using the infinite–finite range inequality Theorem 2.1(d),

$$\|PW\|_{L_\infty(\mathbb{R})} \le C_3 \frac{n}{a_n} \int_{-a_n}^{a_n} |PW|(t) \, dt. \tag{2.22}$$

Next, let

$$v(t) := (1-t^2)^{-1/2}, \qquad t \in (-1,1),$$

be the Chebyshev weight, and let

$$K_n(x,t) := \sum_{j=0}^{n-1} p_j(v,x)p_j(v,t)$$

be the corresponding kernel function. It is well known [12, p. 108] that

$$K_n(x,x) \sim n, \qquad |x| \le 1; \tag{2.23}$$

$$|K_n(x,t)| \le C_4 \min\left\{n, \frac{1}{|x-t|}\right\}, \qquad x,\ t \in [-1,1]. \tag{2.24}$$

Next, let $\hat{P}_{a_{2Ln}}$ be the polynomial of Lemma 2.5 of degree $O(a_{2Ln}\log a_{2Ln}) = o(n)$ by (2.14) (recall $A > 1$). We now fix x s.t. $|x| \le a_{\sigma n}$ and apply (2.22) to the polynomial

$$R(t) := P(t)\hat{P}_{a_{2Ln}}(t)K_n^2\left(\frac{x}{a_{2Ln}}, \frac{t}{a_{2Ln}}\right),$$

where $P \in \mathcal{P}_{Ln}$. Note that R has degree at most $Ln + o(n) + 2n \le 2Ln$, for $n \ge n_1$, say. By (2.22),

$$\left|P(x)\hat{P}_{a_{2Ln}}(x)K_n^2\left(\frac{x}{a_{2Ln}}, \frac{x}{a_{2Ln}}\right)\right| W(x) = |RW|(x)$$
$$\le C_3\frac{2Ln}{a_{2Ln}}\int_{-a_{2Ln}}^{a_{2Ln}} |PW|(t)|\hat{P}_{a_{2Ln}}(t)|K_n^2\left(\frac{x}{a_{2Ln}}, \frac{t}{a_{2Ln}}\right) dt.$$

By Lemma 2.5 and (2.23), we obtain for $|x| \le a_{\sigma n} (< a_{2Ln})$

$$|PW|(x)(1+x^2)^\nu \le C_4\frac{1}{na_n}\int_{-a_{2Ln}}^{a_{2Ln}} |PW|(t)(1+t^2)^\nu K_n^2\left(\frac{x}{a_{2Ln}}, \frac{t}{a_{2Ln}}\right) dt. \tag{2.25}$$

Then by Theorem 2.1(a) and as $\sigma < 1$, we have

$$\begin{aligned}
&\sum_{|x_{jn}|\le\sigma a_n} \lambda_{jn}|P(x_{jn})|W^{-1}(x_{jn})(1+x_{jn}^2)^\nu \\
&\qquad \le C_5\frac{a_n}{n}\sum_{|x_{jn}|\le\sigma a_n} |PW|(x_{jn})(1+x_{jn}^2)^\nu \\
&\qquad \le C_5\int_{-a_{2Ln}}^{a_{2Ln}} |PW|(t)(1+t^2)^\nu T(t)\,dt,
\end{aligned} \tag{2.26}$$

where by (2.25),

$$\begin{aligned}
T(t) &:= n^{-2}\sum_{|x_{jn}|\le\sigma a_n} K_n^2\left(\frac{x_{jn}}{a_{2Ln}}, \frac{t}{a_{2Ln}}\right) \\
&\le C_6 n^{-2}\sum_{|x_{jn}|\le\sigma a_n} \min\left\{n, \frac{a_{2Ln}}{|x_{jn}-t|}\right\}^2,
\end{aligned}$$

by (2.24). Now for $|x_{jn}| \le \sigma a_n$, we have (see Theorem 2.1(b)) uniformly in j and n,

$$x_{j-1,n} - x_{j+1,n} \sim \frac{a_n}{n}.$$

It follows that by reordering the x_{jn} in terms of increasing distance from t, and splitting into sums involving even and odd j, we can bound $T(t)$, uniformly for $t \in \mathbb{R}$, by

$$T(t) \le C_7 n^{-2} \left\{ n^2 + \sum_{j=1}^{n} \left(\frac{a_{2Ln}}{j a_n / n} \right)^2 \right\}$$
$$\le C_8 \left\{ 1 + \sum_{j=1}^{n} j^{-2} \right\} \le C_9.$$

Here we have used $a_{2Ln}/a_n = O(1)$. Substituting into (2.26) yields the lemma. □

Our last quadrature estimate in this section is essentially an estimate for part of the Lebesgue function of Lagrange interpolation.

LEMMA 2.7. *Let $\beta \in (0,2), \nu \in \mathbb{R}$, and*

$$\sum(x) := \sum_{K\ :|x_{kn}| \ge \beta a_n} |\ell_{kn}(x)| W^{-1}(x_{kn})(1+|x_{kn}|)^{-\nu}. \tag{2.27}$$

Then

$$W(x) \sum(x) \le C a_n^{-\nu} \begin{cases} 1, & |x| \le \beta a_n/2, \\ a_n^{1/2} |p_n W|(x) + \log n, & \beta a_n/2 \le |x| \le 2a_n, \\ a_n/|x|, & |x| \ge 2a_n. \end{cases} \tag{2.28}$$

Proof. We first observe that $|x_{kn}| \sim a_n$ uniformly for $|x| \ge \beta a_n$, so it suffices to consider the case $\nu = 0$. Next, with the definition (2.16), we note that (2.5) persists for $j = 1$ and n. So

$$\begin{aligned} &\frac{a_n^{3/2}}{n} \sum_{|x_{kn}| \ge \beta a_n} (\max\{n^{-2/3}, 1 - |x_{kn}|/a_n\})^{-1/4} \\ &\quad \sim a_n^{1/2} \sum_{|x_{kn}| \ge \beta a_n} (x_{k-1,n} - x_{k+1,n})(\max\{n^{-2/3}, 1 - |x_{kn}|/a_n\})^{1/4} \\ &\quad \le a_n^{1/2} \sum_{|x_{kn}| \ge \beta a_n} (x_{k-1,n} - x_{k+1,n}) \\ &\quad \le a_n^{1/2} 2 x_{1n} \le C_1 a_n^{3/2} \end{aligned} \tag{2.29}$$

by (2.4). We now consider three ranges of x, and use (2.12) for x_{jn} close to x, and (2.11) for x_{jn} not close to x.

Case I. $|x| \le \beta a_n/2$. For this range of x, we have

$$|x_{kn} - x| \sim |x_{kn}| \sim a_n.$$

Then (2.11) yields

$$\sum(x) \le C_1 \frac{a_n^{3/2}}{n} \sum_{|x_{kn}| \ge \beta a_n} (\max\{n^{-2/3}, 1 - |x_{kn}|/a_n\})^{-1/4} \frac{|p_n(x)|}{a_n}$$
$$\le C_2 a_n^{1/2} |p_n(x)| \le C_3 W^{-1}(x),$$

(recall $\nu = 0$) by (2.29) and the bound (2.6) for p_n, since $\beta/2 < 1$. Hence (2.28).

Case II. $\beta a_n/2 \le |x| \le 2a_n$. This is the most complicated case, and we need to split

$$\begin{aligned}\sum(x) &= \sum_{\substack{|x_{kn}|\ge\beta a_n\\ |x_{kn}|\le\frac12|x|}} + \sum_{\substack{|x_{kn}|\ge\beta a_n\\ \frac12|x|<|x_{kn}|<2|x|}} + \sum_{\substack{|x_{kn}|\ge\beta a_n\\ |x_{kn}|\ge 2|x|}} \\ &=: \sum\nolimits_1(x) + \sum\nolimits_2(x) + \sum\nolimits_3(x).\end{aligned} \tag{2.30}$$

$\sum_1$: For k in $\sum_1$, we have $|x_{kn} - x| \sim |x| \sim a_n$, and so by (2.11),

$$\begin{aligned}\sum\nolimits_1(x) &\le C_4\frac{a_n^{3/2}}{n}\sum_{|x_{kn}|\ge\beta a_n}(\max\{n^{-2/3}, 1-|x_{kn}|/a_n\})^{-1/4}\frac{|p_n(x)|}{a_n} \\ &\le C_5 a_n^{1/2}|p_n(x)|,\end{aligned} \tag{2.31}$$

by (2.29).

$\sum_2$: Choose $\ell = \ell(x)$ such that $x \in [x_{\ell+1,n}, x_{\ell n}]$ and split

$$\sum\nolimits_2(x) := \sum\nolimits_{21}(x) + \sum\nolimits_{22}(x),$$

where $\sum_{21}$ sums over those k in $\sum_2$ for which $k \in [\ell(x)-3, \ell(x)+3]$ and $\sum_{22}$ contains the rest. Here, if $|x| \ge x_{0n}$ the term $\sum_{21}$ is taken as zero. Now by (2.12), we see that

$$\sum\nolimits_{21}(x) \le C_6 W^{-1}(x).$$

Next, if the sum below is over those k considered in $\sum_{22}$, we can use (2.11) and then (2.5) to deduce that

(2.32)

$$\begin{aligned}\sum\nolimits_{22}(x) &\le C_7\frac{a_n^{3/2}}{n}\sum(\max\{n^{-2/3}, 1-|x_{kn}|a_n\})^{-1/4}\frac{|p_n(x)|}{|x-x_{kn}|} \\ &\le C_8 a_n^{1/2}|p_n(x)|\sum\frac{x_{k-1,n}-x_{k+1,n}}{|x-x_{kn}|}(\max\{n^{-2/3}, 1-|x_{kn}|/a_n\})^{1/4} \\ &\le C_9 a_n^{1/2}|p_n(x)|\int_{[-2a_n,2a_n]\setminus[x_{\ell+3,n},x_{\ell-3,n}]}(\max\{n^{-2/3}, 1-|t|/a_n\})^{1/4}\frac{dt}{|x-t|},\end{aligned}$$

with obvious modifications, both here and below, when $|x| \ge x_{0n}$. In the last step we used Theorem 2.1(f) and

$$|x-t| \sim |x - x_{k\pm1,n}| \sim |x - x_{kn}|, \qquad t \in [x_{k+1,n}, x_{k-1,n}],$$

provided $k \notin [\ell(x)-3, \ell(x)+3]$. This is an easy consequence of (2.5):

$$\begin{aligned}\left|\frac{x-t}{x-x_{kn}}\right| &\le \left|\frac{x-x_{k\pm1,n}}{x-x_{kn}}\right| = \left|1+\frac{x_{kn}-x_{k\pm1,n}}{x-x_{kn}}\right| \\ &\le 1+\left|\frac{x_{kn}-x_{k\pm1,n}}{x_{k\pm2,n}-x_{kn}}\right| \le C_{10}\end{aligned}$$

by (2.5). Similarly we may derive a lower bound. We proceed to estimate the integral in the right-hand side of (2.32). Let us suppose, as we may, that $x \geq 0$ and set $\tau =: x/a_n$. Note that for the current range of $x, \tau \in [\beta/2, 2]$. Then we see that the integral in (2.32) equals

$$I := \int_{[-2,2]\setminus[x_{\ell+3,n}/a_n, x_{\ell-3,n}/a_n]} (\max\{n^{-2/3}, 1-|s|\})^{1/4} \frac{dt}{|\tau - s|}.$$

In view of the spacing (2.5), we have

$$x_{\ell+3,n}/a_n \leq \tau - \delta_n; \qquad x_{\ell-3,n}/a_n \geq \tau + \delta_n,$$

where

$$\delta_n = \delta_n(x) := C\frac{1}{n}\max\{n^{-2/3}, 1-|x|/a_n\}^{-1/2},$$

and C is independent of n, x, and hence ℓ. So

$$I \leq 2\int_{[0,2]\setminus[\tau-\delta_n,\tau+\delta_n]} (\max\{n^{-2/3}, 1-s\})^{1/4} \frac{ds}{|\tau - s|}. \tag{2.33}$$

Suppose first that $|1-\tau| \geq n^{-2/3}$ and set $\sigma := \frac{1-\tau}{|1-\tau|} = \pm 1$. Then the substitution $(1-s) = |1-\tau|u$ gives

$$\begin{aligned} I &\leq 2|1-\tau|^{1/4} \int_{[(-1/|1-\tau|),(1/|1-\tau|)]\setminus[\sigma-(\delta_n/|1-\tau|),\sigma+(\delta_n/|1-\tau|)]} \left(\max\left\{\frac{n^{-2/3}}{|1-\tau|}, u\right\}\right)^{1/4} \\ &\quad \times \frac{du}{|\sigma - u|} \\ &\leq C_{11}|1-\tau|^{1/4}\left[\int_{[\sigma-1/2,\sigma-(\delta_n/|1-\tau|)]} \frac{du}{|\sigma-u|} + \int_{[2,(1/|1-\tau|)]} u^{-3/4}\,du\right] \end{aligned}$$

(with appropriate replacements for the lower limits $\sigma - \frac{1}{2}, 2$ if necessary)

$$\begin{aligned} &\leq C_{12}\left[|1-\tau|^{1/4}\log\left(\frac{|1-\tau|}{\delta_n}\right) + 1\right] \\ &\leq C_{13}\left[|1-|x|/a_n|^{1/4}\log(n|1-|x|/a_n|^{3/2}) + 1\right] \\ &\leq C_{14}[|1-|x|/a_n|^{1/4}\log n + 1]. \end{aligned} \tag{2.34}$$

If $|1-\tau| < n^{-2/3}$, then note that $\delta_n = Cn^{-2/3}$. Set $\kappa := n^{2/3}(1-\tau)$ and note that $|\kappa| < 1$. Here we use the substitution $(1-s) = n^{-2/3}u$ in (2.33) to obtain

$$\begin{aligned} I &\leq 2n^{-1/6}\int_{[-n^{2/3},n^{2/3}]/[\kappa-C,\kappa-C]} \max\{1,u\}^{1/4}\frac{du}{|\kappa-u|} \\ &\leq C_{15}n^{-1/6}\int_2^{n^{2/3}} u^{-3/4}\,du \leq C_{16}. \end{aligned}$$

Of course as C is independent of κ, so is C_{16}. So also in this case, we have the estimate (2.34). Summarizing, we have from (2.32) to (2.34) that

$$\begin{aligned} W(x)\sum\nolimits_2(x) &= W(x)\left[\sum\nolimits_{21}(x) + \sum\nolimits_{22}(x)\right] \\ &\leq C_{17}[1 + a_n^{1/2}|p_nW|(x)\{|1-|x|/a_n|^{1/4}\log n + 1\}] \\ &\leq C_{18}[\log n + a_n^{1/2}|p_nW|(x)]. \end{aligned} \tag{2.35}$$

Here we have used the bound (2.6) for the orthonormal polynomials.

$\sum_3$: For this range of k, we have $|x_{kn} - x| \sim |x_{kn}| \sim a_n$, so

$$\begin{aligned}\sum\nolimits_3(x) &\le C_{18}\frac{a_n^{3/2}}{n}\sum_{|x_{kn}|\ge\beta a_n}(\max\{n^{-2/3}, 1-|x_{kn}|/a_n\})^{-1/4}\frac{|p_n(x)|}{a_n} \\ &\le C_{19}a_n^{1/2}|p_n(x)|\end{aligned} \tag{2.36}$$

by (2.29). Substituting the estimates (2.31), (2.35), and (2.36) into (2.30) yields the result in this case.

Case III. $|x| \ge 2a_n$. For this range of x, we have $|x_{kn} - x| \sim |x|$, (recall that $x_{1n} = a_n[1 + o(1)]$) so

$$\begin{aligned}\sum(x) &\le C_{20}\frac{a_n^{3/2}}{n}\sum_{|x_{kn}|\ge\beta a_n}(\max\{n^{-2/3}, 1-|x_{kn}|/a_n\})^{-1/4}\frac{|p_n(x)|}{|x|} \\ &\le C_{21}\frac{a_n^{3/2}}{|x|}|p_n(x)|,\end{aligned}$$

by (2.29). Finally, (2.6) of Theorem 2.1(c) yields

$$|p_nW|(x) \le C_{22}a_n^{-1/2}, \qquad |x| \ge 2a_n. \qquad \square$$

3. Proof of the theorems. In proving Theorem 1.3, we shall split our function into pieces that vanish inside or outside $[-\beta a_n, \beta a_n]$, some $\beta > 0$. The proof requires several preliminary steps. Many of the ideas have been taken from Nevai's fundamental paper [13], though the fact that a_n may grow slower than $n^{1/2}$ forces a completely different approach in the following lemma, and elsewhere. Throughout, W is as in Theorem 1.3, and throughout $1 < p < \infty, \alpha > 0, \Delta \in \mathbb{R}$, and $\hat{\alpha}$ is given by (1.3).

LEMMA 3.1. *Assume that if* $p \le 4$, (1.13) *holds and if* $p > 4$, (1.14) *holds. Let* $\varepsilon > 0, \beta \in (0,2)$ *and assume that* $\{f_n\}_{n=1}^{\infty}$ *is a sequence of functions from* $\mathbb{R}$ *to* $\mathbb{R}$ *such that*

$$f_n(x) = 0, \qquad |x| < \beta a_n, \tag{3.1}$$

and

$$|f_nW|(x) \le \varepsilon(1+|x|)^{-\alpha}, \qquad x \in \mathbb{R}, \quad n \ge 1. \tag{3.2}$$

Then

$$\limsup_{n\to\infty}\|L_n[f_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})} \le C\varepsilon. \tag{3.3}$$

Here C *is independent of* ε, n, *and* $\{f_n\}$.

Proof. Now by Lemma 2.7, and (3.1)–(3.2),

$$\begin{aligned}|L_n[f_n](x)W(x)| &= \left|W(x)\sum_{|x_{kn}|\ge\beta a_n}\ell_{kn}(x)f_n(x_{kn})\right| \\ &\le \varepsilon W(x)\sum_{|x_{kn}|\ge\beta a_n}|\ell_{kn}(x)|W^{-1}(x_{kn})(1+|x_{kn}|)^{-\alpha} \\ &\le C_1\varepsilon a_n^{-\alpha}\begin{cases}1, & |x| \le \beta a_n/2, \\ a_n^{1/2}|p_nW|(x) + \log n, & \beta a_n/2 \le |x| \le 2a_n, \\ a_n/|x|, & |x| \ge 2a_n.\end{cases}\end{aligned} \tag{3.4}$$

Then

$$\begin{aligned}\tau_n^{(1)} &:= \|L_n[f_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\le \beta a_n/2)}\\ &\le C_1\varepsilon a_n^{-\alpha}\|(1+|x|)^{-\Delta}\|_{L_p(|x|\le\beta a_n/2)}\\ &\le C_2\varepsilon a_n^{-\alpha}\begin{cases}1, & \Delta p>1,\\ (\log n)^{1/p}, & \Delta p=1,\\ a_n^{1/p-\Delta}, & \Delta p<1.\end{cases}\end{aligned}$$

Here, if $\Delta p \ge 1$, this term is $o(1)$ as $\alpha > 0$ and a_n grows faster than some positive power of n. Suppose now that $\Delta p < 1$. Note that if $p > 4$, then the power of n in (1.14) is positive, so the power of a_n there, namely $1/p - (\hat{\alpha} + \Delta)$, must be negative. Hence (1.13) holds for all $p > 1$. (We shall use this repeatedly.) So if $\Delta p < 1$, (1.13) shows that

$$a_n^{-\alpha+1/p-\Delta} \le a_n^{-\hat{\alpha}+1/p-\Delta} = o(1). \tag{3.5}$$

Hence in all cases

$$\lim_{n\to\infty} \tau_n^{(1)} = 0. \tag{3.6}$$

Next, from (3.4),

$$\begin{aligned}\tau_n^{(2)} &:= \|L_n[f_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\beta a_n/2\le|x|\le 2a_n)}\\ &\le C_2\varepsilon a_n^{-\alpha}\{a_n^{1/2-\Delta}\|p_nW\|_{L_p[\beta a_n/2,2a_n]} + (\log n)a_n^{1/p-\Delta}\}\\ &\le C_3\varepsilon a_n^{-(\alpha+\Delta)+1/p}\begin{cases}1, & p<4\\ (\log n)^{1/4}, & p=4\\ n^{(1/6)(1-p/4)}, & p>4\end{cases} + C_3\varepsilon(\log n)a_n^{-(\alpha+\Delta)+1/p},\end{aligned}$$

by Theorem 2.2(a). Recall that for all $p > 1$, $-(\alpha + \Delta) + 1/p < 0$ (see (3.5)) and a_n is of polynomial growth, so the second term in the last right-hand side is always $o(1)$. By the same token, for $p \le 4$, this entire last right-hand side is $o(1)$. When $p > 4$ and $\alpha > 1$, then $\alpha > \hat{\alpha}$ (see (1.3)) so (1.14) shows that this whole last right-hand side is $o(1)$. When $p > 4$ and $\alpha \le 1$, our assumption (1.14) shows that this last term is $O(\varepsilon)$. So in all cases,

$$\limsup_{n\to\infty} \tau_n^{(2)} \le C_4\varepsilon. \tag{3.7}$$

Finally, from (3.4),

$$\begin{aligned}\tau_n^{(3)} &:= \|L_n[f_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\ge 2a_n)}\\ &\le C_5 a_n^{-\alpha+1}\||x|^{-1}(1+|x|)^{-\Delta}\|_{L_p(|x|\ge 2a_n)}.\end{aligned}$$

Now from (1.13),

$$\Delta > \frac{1}{p} - \hat{\alpha} \ge \frac{1}{p} - 1,$$

so

$$p(\Delta + 1) > 1, \tag{3.8}$$

and hence

$$\tau_n^{(3)} \le C_6 a_n^{-\alpha+1} a_n^{1/p-(1+\Delta)} = o(1),$$

by (3.5). Together with (3.6) and (3.7), this yields the result. □

Note that we needed only the weaker (1.14) and not (1.15) in the above lemma. Having dealt with functions that vanish inside $(-\beta a_n, \beta a_n)$, we turn to functions that vanish outside this interval. First we estimate the L_p norms of such functions outside $[-2\beta a_n, 2\beta a_n]$ and in Lemma 3.4 below, we shall use Hilbert transforms to estimate their L_p norms over $[-2\beta a_n, 2\beta a_n]$.

LEMMA 3.2. *Assume that if* $p \le 4$, (1.13) *holds; if* $p > 4$ *and* $\alpha \ne 1$, (1.14) *holds; and if* $p > 4$ *and* $\alpha = 1$, (1.15) *holds. Let* $\varepsilon > 0, \beta \in (0,1)$ *and assume that* $\{\psi_n\}_{n=1}^{\infty}$ *is a sequence of functions from* $\mathbb{R}$ *to* $\mathbb{R}$ *such that*

$$\psi_n(x) = 0, \qquad |x| \ge \beta a_n, \tag{3.9}$$

and

$$|\psi_n(W)|(x) \le \varepsilon(1+|x|)^{-\alpha}, \qquad x \in \mathbb{R}, \quad n \ge 1. \tag{3.10}$$

Then

$$\limsup_{n\to\infty} \|L_n[\psi_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\ge 2\beta a_n)} \le C\varepsilon. \tag{3.11}$$

Here C *is independent of* $\varepsilon, n,$ *and* $\{\psi_n\}$.

Proof. First note that for $|x| \ge 2\beta a_n$ and $|x_{kn}| \le \beta a_n, |x_{kn} - x| \sim |x|$. Then, from Theorem 2.2(b), we have for $|x| \ge 2\beta a_n$

$$\begin{aligned}
|L_n[\psi_n](x)| &= \left| \sum_{|x_{kn}|<\beta a_n} \ell_{kn}(x)\psi_n(x_{kn}) \right| \\
&\le \varepsilon \sum_{|x_{kn}|<\beta a_n} |\ell_{kn}(x)| W^{-1}(x_{kn})(1+|x_{kn}|)^{-\alpha} \\
&\le C_1 \varepsilon \frac{a_n^{3/2}}{n} \frac{|p_n(x)|}{|x|} \sum_{|x_{kn}|<\beta a_n} (1+|x_{kn}|)^{-\alpha} \\
&\le C_2 \varepsilon a_n^{1/2} \frac{|p_n(x)|}{|x|} \sum_{|x_{kn}|<\beta a_n} (x_{k-1,n} - x_{k+1,n})(1+|x_{kn}|)^{-\alpha} \qquad \text{(by (2.5))} \\
&\le C_3 \varepsilon a_n^{1/2} \frac{|p_n(x)|}{|x|} \int_{-2\beta a_n}^{2\beta a_n} (1+|t|)^{-\alpha}\, dt;
\end{aligned}$$

recall that $\beta < 1$ and see (2.18). Let

$$(\log n)^* := \begin{cases} \log n, & \alpha = 1, \\ 1, & \text{otherwise.} \end{cases}$$

We see (by examining $\alpha <, =, > 1$) that

$$\int_{-2\beta a_n}^{2\beta a_n} (1+|t|)^{-\alpha}\, dt \le C_4 a_n^{1-\hat{\alpha}} (\log n)^*, \qquad n \ge 2.$$

Then

$$\begin{aligned}
&\|L_n[\psi_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\geq 2\beta a_n)} \\
&\quad \leq C_5 \varepsilon a_n^{3/2-\hat{\alpha}} (\log n)^* \|p_n(x)W(x)|x|^{-1-\Delta}\|_{L_p(|x|\geq 2\beta a_n)} \\
&\quad \leq C_5 \varepsilon a_n^{1/2-\hat{\alpha}-\Delta} (\log n)^* \|p_n(x)W(x)\|_{L_p(\mathbb{R})} \quad \text{(recall } \Delta+1>0\text{; see (3.8))} \\
&\quad \leq C_6 \varepsilon a_n^{1/p-(\hat{\alpha}+\Delta)} (\log n)^* \times \begin{cases} 1, & p<4, \\ (\log n)^{1/4}, & p=4, \\ n^{(1/6)(1-4/p)}, & p>4, \end{cases}
\end{aligned}$$

by Theorem 2.2(a). Now if $p \leq 4$, we know that (1.13) holds and a_n is of polynomial growth, so the last right-hand side is $o(1)$. If $p > 4$ and $\alpha \neq 1$, our hypothesis (1.14) ensures that we obtain $O(\varepsilon)$; if $p > 4$ and $\alpha = 1$, our hypothesis (1.15) ensures that again we obtain $O(\varepsilon)$. □

Before proceeding to our third lemma on the L_p norms of Lagrange interpolants, we need the Hilbert transform $H[f](x)$, and its boundedness in suitable weighted L_p spaces. Recall that

$$H[f]x := \lim_{\varepsilon \to 0+} \int_{|t-x|\geq \varepsilon} \frac{f(t)}{t-x}\, dt,$$

and if $f \in L_1(\mathbb{R})$, then this limit exists a.e.

LEMMA 3.3. *Let $1 < p < \infty, s < 1 - 1/p; S > -1/p; s \leq S$. Then for measurable $f : \mathbb{R} \to \mathbb{R}$, for which the right-hand side is finite,*

$$\|H[f](x)(1+|x|)^s\|_{L_p(\mathbb{R})} \leq C\|f(x)(1+|x|)^S\|_{L_p(\mathbb{R})}. \tag{3.12}$$

Proof. This is the special case $R = r = 0$ of Lemma 8 in [11, p. 440]. □

LEMMA 3.4. *Assume that (1.13) holds. Let $\varepsilon > 0, \beta \in (0, 1/2)$, and assume that $\{\psi_n\}_{n=1}^{\infty}$ is a sequence of functions from $\mathbb{R}$ to $\mathbb{R}$ such that*

$$\psi_n(x) = 0, \qquad |x| \geq \beta a_n, \tag{3.13}$$

and

$$|\psi_n W|(x) \leq \varepsilon(1+|x|)^{-\alpha}, \qquad x \in \mathbb{R}, \quad n \geq 1. \tag{3.14}$$

Then

$$\limsup_{n\to\infty} \|L_n[\psi_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\leq 2\beta a_n)} \leq C\varepsilon. \tag{3.15}$$

Here C is independent of ε, n, and $\{\psi_n\}$.

Proof. Define

$$G(x) := W(x)^{-1}(1+|x|)^{-\alpha}, \qquad x \in \mathbb{R}.$$

Furthermore, for $f : \mathbb{R} \to \mathbb{R}$ such that $fW \in L_1(\mathbb{R})$, we let $S_n[f](x)$ denote the nth partial sum of the orthonormal expansion of f in $\{p_j\}_0^{\infty}$ (so that $S_n[f](x)$ is a linear combination of $p_0, p_1, \ldots, p_{n-1}$). We first show that

$$\begin{aligned}
&\|L_n[\psi_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\leq 2\beta a_n)} \\
&\qquad \leq C_2\varepsilon \sup_{\|h\|_{L_\infty(\mathbb{R})}\leq 1} \|S_n[hG](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\leq 2\beta a_n)}.
\end{aligned} \tag{3.16}$$

Let χ_n denote the characteristic function of $[-2\beta a_n, 2\beta a_n]$, and for $n \geq 1$, let

$$g_n(x) := \operatorname{sign}\{L_n[\psi_n](x)\}|L_n[\psi_n](x)|^{p-1}\chi_n(x)W^{p-2}(x)(1+|x|)^{-\Delta p}.$$

Observe that

$$\begin{aligned}
&\|L_n[\psi_n](x)W(x)(1+|x|)^{-\Delta}\|^p_{L_p(|x|\leq 2\beta a_n)}\\
&\quad = \int_{-\infty}^{\infty} L_n[\psi_n](x)g_n(x)W^2(x)\,dx\\
&\quad = \int_{-\infty}^{\infty} L_n[\psi_n](x)S_n[g_n](x)W^2(x)\,dx\\
&\quad \text{(orthogonality of } g - S_n[g] \text{ to } \mathcal{P}_{n-1})\\
&\quad = \sum_{k=1}^{n} \lambda_n\psi_n(x_{kn})S_n[g_n](x_{kn}) \qquad \text{(Gauss quadrature)}\\
&\quad \leq \epsilon \sum_{|x_{kn}|<\beta a_n} \lambda_{kn}|S_n[g_n](x_{kn})|W^{-1}(x_{kn})(1+|x_{kn}|)^{-\alpha} \qquad \text{(by (3.14))}\\
&\quad \leq C_3\varepsilon \int_{-\infty}^{\infty} |S_n[g_n](x)|W(x)(1+|x|)^{-\alpha}\,dx\\
&\quad \text{(by Lemma 2.6 and as } (1+|t|)^{-\alpha} \sim (1+t^2)^{-\alpha/2})\\
&\quad = C_3\varepsilon \int_{-\infty}^{\infty} S_n[g_n](x)h_n(x)G(x)W^2(x)\,dx,
\end{aligned}$$

where we have set $h_n(x) := \operatorname{sign}(S_n[g_n](x))$, and we have used the definition of G. Then orthogonality and Hölder's inequality (with $q = p/(p-1)$) show that this last expression can be continued as

$$\begin{aligned}
&C_3\varepsilon \int_{-\infty}^{\infty} g_n(x)S_n[h_nG](x)W^2(x)\,dx\\
&\quad = C_3\varepsilon \int_{|x|\leq 2\beta a_n} g_n(x)S_n[h_nG](x)W^2(x)\,dx\\
&\quad \leq C_3\varepsilon \|g_n(x)W(x)(1+|x|)^{\Delta}\|_{L_q(|x|\leq 2\beta a_n)}\\
&\quad \times \|S_n[h_nG](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\leq 2\beta a_n)}\\
&\quad = C_3\varepsilon \|L_n[\psi_n](x)W(x)(1+|x|)^{-\Delta}\|^{p-1}_{L_p(|x|\leq 2\beta a_n)}\\
&\quad \times \|S_n[h_nG](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x|\leq 2\beta a_n)}.
\end{aligned}$$

In the last right-hand side we have used the definition of g_n. Cancelling the $(p-1)$th power of the norm in the last right-hand side, and in the left-hand side after the definition of g_n gives (3.16). Next, it is a well-known consequence of the Christoffel–Darboux formula that for $f : \mathbb{R} \to \mathbb{R}$, for which $fW \in L_1(\mathbb{R})$,

$$S_n[f](x) = \frac{\gamma_{n-1}}{\gamma_n}\{p_n(x)H[p_{n-1}fW^2](x) - p_{n-1}(x)H[p_nfW^2](x)\},$$

where H denotes the Hilbert transform, as above. Then using Theorem 2.1(c) and (e), we have for $|x| \leq 2\beta a_n$ (recall $2\beta < 1$) and $h \in L_\infty(\mathbb{R})$ that

$$|S_n[hG](x)W(x)| \leq C_4 a_n^{1/2} \sum_{j=n-1}^{n} |H[p_jhGW^2](x)|. \tag{3.17}$$

Choose $\sigma \in (2\beta, 1)$, and let χ_n^* be the characteristic function of $[-\sigma a_n, \sigma a_n], n \geq 1$. Moreover, let $\|h\|_{L_\infty(\mathbb{R})} \leq 1$. In estimating the Hilbert transform for $|x| \leq 2\beta a_n$, we write

$$H[p_j hGW^2] = H[p_j(1-\chi_n^*)hGW^2] + H[p_j \chi_n^* hGW^2].$$

Now for $|x| \leq 2\beta a_n$, and $j = n-1, n$,

$$\begin{aligned}
|H[p_j(1-\chi_n^*)hGW^2](x)| &= \left| \int_{|t| \in [\sigma a_n, \infty)} \frac{p_j(t)(hGW^2)(t)}{x-t}\, dt \right| \\
&\leq C_5 \int_{\sigma a_n}^{\infty} |p_j W|(t) t^{-1-\alpha}\, dt \\
&\leq C_5 \left(\int_{\sigma a_n}^{\infty} p_j^2(t) W^2(t)\, dt \right)^{1/2} \left(\int_{\sigma a_n}^{\infty} t^{-2-2\alpha}\, dt \right)^{1/2} \\
&\leq C_6 a_n^{-1/2-\alpha}.
\end{aligned}$$

This last estimate and (3.17) give for $|x| \leq 2\beta a_n$,

$$|S_n[hG](x)W(x)| \leq C_4 a_n^{1/2} \sum_{j=n-1}^{n} |H[p_j \chi_n^* hGW^2](x)| + C_7 a_n^{-\alpha}. \tag{3.18}$$

Now let us set for some small $\delta > 0$,

$$\hat{\Delta} := \min\{\Delta, 1/p - \delta\}.$$

We now use Lemma 3.3 with $s = S = -\hat{\Delta}$. Note that

$$S = -\hat{\Delta} = \max\{-\Delta, \delta - 1/p\} > -1/p;$$

also, (1.13) (which, the reader may recall, holds for all $p > 1$) gives

$$-\Delta < -1/p + \hat{\alpha} \leq -1/p + 1,$$

so if $\delta < 1$, then

$$s = -\hat{\Delta} = \max\{-\Delta, \delta - 1/p\} < 1 - 1/p.$$

Thus the requirements of Lemma 3.3 are met. Then as $\hat{\Delta} \leq \Delta$, (3.18), and Lemma 3.3 yield

$$\begin{aligned}
&\|S_n[hG](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(|x| \leq 2\beta a_n)} \\
&\quad \leq C_4 a_n^{1/2} \sum_{j=n-1}^{n} \|H[p_j \chi_n^* hGW^2](x)(1+|x|)^{-\hat{\Delta}}\|_{L_p(\mathbb{R})} \\
&\qquad + C_7 a_n^{-\alpha} \|(1+|x|)^{-\Delta}\|_{L_p(|x| \leq 2\beta a_n)} \\
&\quad \leq C_8 a_n^{1/2} \sum_{j=n-1}^{n} \|(p_j \chi_n^* hGW^2)(x)(1+|x|)^{-\hat{\Delta}}\|_{L_p(\mathbb{R})} \\
&\qquad + C_8 a_n^{-\alpha} \begin{cases} 1, & \Delta p > 1, \\ (\log n)^{1/p}, & \Delta p = 1, \\ a_n^{1/p - \Delta}, & \Delta p < 1, \end{cases} \\
&\quad \leq C_8 a_n^{1/2} \sum_{j=n-1}^{n} \|(p_j W)(x)(1+|x|)^{-(\alpha+\hat{\Delta})}\|_{L_p(|x| \leq a_{\sigma n})} + o(1),
\end{aligned} \tag{3.19}$$

since $\alpha > 0$, and a_n is of polynomial growth, while also by (1.13)

$$-\alpha + 1/p - \Delta \le -\hat{\alpha} + 1/p - \Delta < 0. \tag{3.20}$$

Then using the bound (2.6) (recall $\sigma < 1$), we can continue (3.19) as

$$\begin{aligned} &\le C_9 \|(1+|x|)^{-(\alpha+\hat{\Delta})}\|_{L_p(|x|\le a_{\sigma n})} + o(1) \\ &\le C_{10}, \end{aligned}$$

as again (3.20) implies that

$$p(\alpha + \hat{\Delta}) > 1. \tag{3.21}$$

Finally, this bound and (3.16) yield the result. □

Proof of the sufficient conditions of Theorem 1.3. We remark first that the sufficient conditions in Theorem 1.3 imply those in Lemmas 3.1, 3.2, and 3.4. (Recall here that (1.15) implies (1.14).) Let $\varepsilon > 0$. We can find a polynomial P such that

$$|f - P|(x)W(x)(1+|x|)^{\alpha} \le \varepsilon, \qquad x \in \mathbb{R}. \tag{3.22}$$

(Cf. [2, p. 180].) Then for n large enough,

(3.23)

$$\begin{aligned} &\|(f - L_n[f])(x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})} \\ &\quad \le \|(f-P)(x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})} + \|L_n[P-f](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})} \\ &\quad \le \varepsilon \|(1+|x|)^{-(\alpha+\Delta)}\|_{L_p(\mathbb{R})} + \|L_n[P-f](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})}. \end{aligned}$$

Here $p(\alpha+\Delta) \ge p(\hat{\alpha}+\Delta) > 1$ (see (3.21)), so the first norm in this last right-hand side is finite. Next, let χ_n denote the characteristic function of $[-a_n/4, a_n/4]$ and write

$$P - f = (P-f)\chi_n + (P-f)(1-\chi_n) =: \psi_n + f_n.$$

We can apply Lemma 3.1 to $\{f_n\}_{n=1}^{\infty}$ with $\beta = 1/4$ to deduce that

$$\limsup_{n\to\infty} \|L_n[f_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})} \le C_1 \varepsilon.$$

Next, in view of (3.22), $\{\psi_n\}_{n=1}^{\infty}$ satisfy (3.9)–(3.10) and (3.13)–(3.14) in Lemma 3.2 and 3.4 with $\beta = 1/4$, so those lemmas yield

$$\limsup_{n\to\infty} \|L_n[\psi_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})} \le C_2 \varepsilon.$$

So

$$\limsup_{n\to\infty} \|L_n[P-f](x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})} \le C_3 \varepsilon. \tag{3.24}$$

Combining (3.23) and (3.24), we have

$$\limsup_{n\to\infty} \|(f - L_n[f])(x)W(x)(1+|x|)^{-\Delta}\|_{L_p(\mathbb{R})} \le C_4 \varepsilon.$$

Letting $\varepsilon \to 0+$ yields the result. □

In the converse direction, we need the following lemma.

LEMMA 3.5. *Let $\sigma \in (0,1), \eta \in (0, 1-\sigma), 1 < p < \infty$. Then there exists C such that for $n \geq 1$ and P of degree at most ηn, we have*

$$\|P\|_{L_p[-a_{\sigma n}, a_{\sigma n}]} \leq C a_n^{1/2} \sum_{j=n-1}^{n} \|p_j W P\|_{L_p[-a_n, a_n]}. \tag{3.25}$$

Proof. Choose $\sigma' \in (\sigma, 1-\eta)$. Let $\delta \in (0,1)$ be such that $\delta\sigma' > \sigma$. For m large enough, Theorem 2.1(a) shows that

$$\lambda_m^{-1}(W,x) \sim \frac{m}{a_m} W^{-1}(x), |x| \leq a_{2\delta m}.$$

Moreover, we can replace $\sim$ by $\leq C \times$ for all $x \in \mathbb{R}$. (We have replaced W by $W^{1/2}$ in (2.3) and used also $a_m(W^{1/2}) = a_{2m}(W)$. Of course our a_ℓ above is $a_\ell(W)$.) Applying this with $m := [\sigma' n/2]$ and using

$$2\delta m = 2\delta[\sigma' n/2] > \sigma n,$$

with n large enough, yields

$$\lambda_{[\sigma' n/2]}^{-1}(W,x) W(x) \begin{cases} \sim n/a_n, & |x| \leq a_{\sigma n}, \\ \leq Cn/a_n, & x \in \mathbb{R}. \end{cases} \tag{3.26}$$

Let P be of degree $\leq \eta n$, and

$$R(x) := \frac{a_n}{n} P(x) \lambda_{[\sigma' n/2]}^{-1}(W,x). \tag{3.27}$$

Then R has degree $\leq \eta n + 2[\sigma' n/2] \leq (\eta + \sigma')n < n-1$, for n large enough, by choice of σ'. Next, with the definition for S_n in Lemma 3.4,

$$R(x) = S_n[R](x) = \frac{\gamma_{n-1}}{\gamma_n} \{p_n(x) H[R p_{n-1} W^2](x) - p_{n-1}(x) H[R p_n W^2](x)\},$$

by the Christoffel–Darboux formula. Using our bounds in Theorem 2.1, we obtain for $|x| \leq a_{\sigma n}$,

$$|RW|(x) \leq C_1 a_n^{1/2} \sum_{j=n-1}^{n} |H[R p_j W^2](x)|.$$

Using Riesz's theorem that H is a bounded operator from $L_p(\mathbb{R})$ to $L_p(\mathbb{R})$ (the special case $S = s = 0$ of Lemma 3.3), we obtain

$$\|RW\|_{L_p[-a_{\sigma n}, a_{\sigma n}]} \leq C_2 a_n^{1/2} \sum_{j=n-1}^{n} \|R p_j W^2\|_{L_p[-a_n, a_n]},$$

where we also used the infinite–finite range inequality (2.8) for the weight W^2. Now (3.26) shows that

$$|RW|(x) \begin{cases} \sim |P(x)|, & |x| \leq a_{\sigma n}, \\ \leq C|P(x)|, & x \in \mathbb{R}. \end{cases}$$

Then (3.25) follows. □

Proof of the necessary conditions of Theorem 1.3. Let $\Delta \in \mathbb{R}, \alpha > 0$. We assume the convergence (1.11) for every continuous $f : \mathbb{R} \to \mathbb{R}$ satisfying (1.12). Let $\eta(x) : \mathbb{R} \to (0, \infty)$ be an even continuous function that is decreasing in $[0, \infty)$, with

$$\eta(x) \geq (\log(2 + |x|))^{-1/(2p)}, \qquad x \in [0, \infty), \tag{3.28}$$

but

$$\lim_{x \to \infty} \eta(x) = 0. \tag{3.29}$$

We let X be the space of all continuous functions $f : \mathbb{R} \to \mathbb{R}$ with

$$\|f\|_X := \max_{x \in \mathbb{R}} |f(x)| W(x)(1 + |x|)^{\alpha} \eta(x)^{-1} < \infty.$$

Furthermore, let Y be the space of all measurable functions $f : \mathbb{R} \to \mathbb{R}$ with

$$\|f\|_Y := \|(fW)(x)(1 + |x|)^{-\Delta}\|_{L_p(\mathbb{R})} < \infty.$$

Now each $f \in X$ satisfies (1.12), so our hypothesis ensures that

$$\lim_{n \to \infty} \|f - L_n[f]\|_Y = 0.$$

By the uniform boundedness principle (recall that X is a Banach space), there exists $C_1 > 0$ such that

$$\|f - L_n[f]\|_Y \leq C_1 \|f\|_X \quad \forall n \geq 1, \quad \forall f \in X.$$

In particular, as $L_1[f] = f(0)$ (recall $p_n(W^2, x) = x$) we obtain for every continuous $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = 0$, that

$$\|f\|_Y \leq C_1 \|f\|_X,$$

provided the norm on the right-hand side is finite. Hence, for every $n \geq 1$ and every continuous f with $f(0) = 0$ for which the right-hand side is finite,

$$\|L_n[f](x) W(x)(1 + |x|)^{-\Delta}\|_{L_p(\mathbb{R})} \leq C_2 \|(fW)(x)(1 + |x|)^{\alpha} \eta(x)^{-1}\|_{L_\infty(\mathbb{R})}. \tag{3.30}$$

Now choose $g_n, n \geq 1$, such that g_n is continuous in $\mathbb{R}$, $g_n = 0$ in $[0, \infty) \cup (-\infty, -a_{n/2})$;

$$\|g_n\|_X = \|(g_n W)(x)(1 + |x|)^{\alpha} \eta(x)^{-1}\|_{L_\infty(\mathbb{R})} = 1; \tag{3.31}$$

and for $x_{jn} \in [-a_{n/2}, 0)$,

$$(g_n W)(x_{jn})(1 + |x_{jn}|)^{\alpha} \eta(x_{jn})^{-1} \operatorname{sign}(p_n'(x_{jn})) = 1.$$

Then for $x > 0$, our choice of g_n and then (2.11) yield

$$\begin{aligned}
|L_n[g_n](x)| &= \left| \sum_{x_{jn} \in [-a_{n/2}, 0)} g_n(x_{jn}) \frac{p_n(x)}{p_n'(x_{jn})(x - x_{jn})} \right| \\
&= \sum_{x_{jn} \in [-a_{n/2}, 0)} \left| \frac{p_n(x)}{p_n'(x_{jn})(x - x_{jn})} \right| W^{-1}(x_{jn})(1 + |x_{jn}|)^{-\alpha} \eta(x_{jn}) \\
&= \sum_{x_{jn} \in [-a_{n/2}, 0)} |\ell_{jn}(x)| W^{-1}(x_{jn})(1 + |x_{jn}|)^{-\alpha} \eta(x_{jn}) \\
&\geq C_3 \frac{a_n^{3/2}}{n} \sum_{x_{jn} \in [-a_{n/2}, 0)} \frac{|p_n(x)|}{x + |x_{jn}|} (1 + |x_{jn}|)^{-\alpha} \eta(x_{jn}) \\
&\geq C_4 \eta(a_n) a_n^{1/2} |p_n(x)| \sum_{x_{jn} \in [-a_{n/2}, 0)} \frac{x_{j-1,n} - x_{j+1,n}}{x + |x_{jn}|} (1 + |x_{jn}|)^{-\alpha}
\end{aligned}$$

by (2.5). At least for $x \geq 1$, we have (see (2.18))

$$x + |x_{jn}| \sim x + |t|, \qquad t \in (x_{j+1,n}, x_{j-1,n}),$$

so for $x \geq 1$,

$$\begin{aligned}
|L_n[g_n](x)| &\geq C_5 \eta(a_n) a_n^{1/2} |p_n(x)| \int_0^{a_{n/4}} \frac{(1+t)^{-\alpha}}{x+t}\, dt \\
&\geq C_5 \eta(a_n) a_n^{1/2} |p_n(x)| \int_0^{\min\{a_{n/4}, x\}} \frac{(1+t)^{-\alpha}}{2x}\, dt \\
&\geq C_6 \eta(a_n) a_n^{1/2} \frac{|p_n(x)|}{x} \begin{cases} 1, & \alpha > 1, \\ \log(1 + \min\{a_{n/4}, x\}), & \alpha = 1, \\ (\min\{a_{n/4}, x\})^{1-\alpha}, & \alpha < 1. \end{cases}
\end{aligned}$$

Let us set

$$(\log x)^* := \begin{cases} \log(1+x) & \text{if } \alpha = 1, \\ 1 & \text{otherwise.} \end{cases}$$

By considering $x \in [1, a_{n/4}]$ and $x \in [a_{n/4}, 2a_n]$ separately, we see that we can rewrite the above as

$$|L_n[g_n](x)| \geq C_7 \eta(a_n) a_n^{1/2} |p_n(x)| x^{-\hat{\alpha}} (\log x)^* \tag{3.32}$$

for $x \in [1, 2a_n]$ (recall the definition (1.3) of $\hat{\alpha}$). Then we obtain from (3.28) that for $x \in [1, 2a_n]$,

$$|L_n[g_n](x)| \geq C_8 (\log n)^{-1/(2p)} a_n^{1/2} |p_n(x)| x^{-\hat{\alpha}}.$$

Hence, using (3.30)–(3.31),

$$\begin{aligned}
C_2 &\geq \|L_n[g_n](x) W(x) (1+|x|)^{-\Delta}\|_{L_p[1,2a_n]} \\
&\geq C_9 (\log n)^{-1/(2p)} a_n^{1/2} \|(p_n W)(x)(1+|x|)^{-(\hat{\alpha}+\Delta)}\|_{L_p[1,2a_n]} \\
&\geq C_{10} (\log n)^{-1/(2p)} a_n^{1/2} \|(p_n W)(x)(1+|x|)^{-(\hat{\alpha}+\Delta)}\|_{L_p[0,2a_n]} - C_{11} (\log n)^{-1/(2p)},
\end{aligned}$$

where we have used the bound (2.6) for p_n in [0, 1]. So

$$C_{12} (\log n)^{1/(2p)} \geq a_n^{1/2} \|(p_n W)(x)(1+|x|)^{-(\hat{\alpha}+\Delta)}\|_{L_p[-2a_n, 2a_n]}.$$

Now let $\hat{P}_{2a_n}$ be the polynomial of Lemma 2.5 of degree $O(a_n \log a_n) = o(n)$ such that for $|x| \leq 2a_n$,

$$\hat{P}_{2a_n}(x) \sim (1+x^2)^{-(\hat{\alpha}+\Delta)/2} \sim (1+|x|)^{-(\hat{\alpha}+\Delta)}.$$

Then we obtain from Lemma 3.5, (with $\sigma = 1/2$, and for example $\eta = 1/4$)

$$\begin{aligned}
C_{13} (\log n)^{1/(2p)} &\geq a_n^{1/2} \sum_{j=n-1}^{n} \|(p_j W)(x) \hat{P}_{2a_n}(x)\|_{L_p[-2a_j, 2a_j]} \\
&\geq C_{14} \|\hat{P}_{2a_n}(x)\|_{L_p[-a_{n/2}, a_{n/2}]} \\
&\geq C_{15} \|(1+|x|)^{-(\hat{\alpha}+\Delta)}\|_{L_p[-a_{n/2}, a_{n/2}]} \\
&\geq C_{16} \begin{cases} a_n^{1/p - (\hat{\alpha}+\Delta)}, & \Delta < 1/p - \hat{\alpha}, \\ (\log n)^{1/p}, & \Delta = 1/p - \hat{\alpha}, \\ 1, & \Delta > 1/p - \hat{\alpha}. \end{cases}
\end{aligned}$$

This is not possible for all large enough n, unless $\Delta > 1/p - \hat{\alpha}$. So we have proved (1.13) for all $1 < p < \infty$ and in particular, for $p \le 4$. This establishes the necessary conditions for $p \le 4$. To prove the necessary conditions for $p > 4$, we return to (3.32). First, for small enough $\delta \in (0,1)$, it is an easy consequence of Theorem 2.1(c) and (d), that

$$\|p_n W\|_{L_p[\delta a_n, 2a_n]} \sim \|p_n W\|_{L_p(\mathbb{R})}.$$

Then by (3.30)–(3.32),

$$\begin{aligned} C_2 &\ge \|L_n[g_n](x)W(x)(1+|x|)^{-\Delta}\|_{L_p[\delta a_n, 2a_n]} \\ &\ge C_{17}\eta(a_n)a_n^{1/2}\|p_n W\|_{L_p[\delta a_n, 2a_n]}a_n^{-(\hat{\alpha}+\Delta)}(\log n)^* \\ &\ge C_{18}\eta(a_n)a_n^{1/p-(\hat{\alpha}+\Delta)}(\log n)^* n^{(1/6)(1-4/p)}, \end{aligned}$$

by (2.10) as $p > 4$. Thus

$$\limsup_{n\to\infty} \eta(a_n)a_n^{1/p-(\hat{\alpha}+\Delta)}(\log n)^* n^{(1/6)(1-4/p)} < \infty, \tag{3.33}$$

for every function η satisfying (3.28)–(3.29). If

$$\limsup_{n\to\infty} a_n^{1/p-(\hat{\alpha}+\Delta)}(\log n)^* n^{(1/6)(1-4/p)} = \infty,$$

then it is easy to construct $\eta(x)$ decreasing slowly enough to 0 to contradict (3.33). So

$$\limsup_{n\to\infty} a_n^{1/p-(\hat{\alpha}+\Delta)}(\log n)^* n^{(1/6)(1-4/p)} < \infty,$$

and it follows that (1.14) is necessary if $\alpha \ne 1$ and (1.15) is necessary if $\alpha = 1$.

Proof of Theorem 1.1. For $W = W_\beta, a_n = Cn^{1/\beta}, n \ge 1$, and the necessary and sufficient conditions in Theorem 1.3 are very easily seen to become those in Theorem 1.1.

Proof of Corollary 1.2. Let

$$\tau(p) := \begin{cases} \frac{1}{p} - \hat{\alpha}, & p \le 4, \\ \frac{1}{p} - \hat{\alpha} + \frac{\beta}{6}\left(1 - \frac{4}{p}\right), & p > 4. \end{cases}$$

For the convergence (1.5) to hold for every $1 < p \le \infty$ and every continuous f satisfying (1.6) with a given $\Delta \in \mathbb{R}$ (independent of p), the necessary conditions of Theorem 1.1 imply that

$$\Delta \ge \tau(p), \qquad 1 < p < \infty.$$

Hence

$$\Delta \ge \lim_{p\to 1+} \tau(p) = 1 - \hat{\alpha};$$

$$\Delta \ge \lim_{p\to\infty} \tau(p) = \frac{\beta}{6} - \hat{\alpha}.$$

So it is necessary that (1.7) holds.

Conversely, for the convergence (1.5) to hold for every $1 < p < \infty$ and every continuous f satisfying (1.6), the sufficient conditions of Theorem 1.1 show that it suffices that

$$\Delta > \tau(p), \qquad 1 < p < \infty. \tag{3.34}$$

First, (1.7) shows that

$$\Delta \geq 1 - \hat{\alpha} > 1/p - \hat{\alpha} = \tau(p), \qquad 1 < p \leq 4.$$

So for $1 < p \leq 4$, (3.34) is fulfilled. Next, we can write for $p > 4$,

$$\tau(p) = \frac{\beta}{6} - \hat{\alpha} + \frac{1}{p}\left(1 - \frac{2\beta}{3}\right).$$

If $1 - (2\beta/3) \geq 0$, then for $4 < p < \infty$

$$\tau(p) \leq \tau(4) = \frac{\beta}{6} - \hat{\alpha} + \frac{1}{4}\left(1 - \frac{2\beta}{3}\right) = -\hat{\alpha} + \frac{1}{4} < 1 - \hat{\alpha} \leq \Delta,$$

by (1.7) so again (3.34) is fulfilled. Finally, if $1 - (2\beta/3) < 0$, then for $4 < p < \infty$

$$\tau(p) < \tau(\infty) = \frac{\beta}{6} - \hat{\alpha} \leq \Delta,$$

by (1.7). So we have shown that (1.7) implies (3.34) for all $1 < p < \infty$, and this establishes that (1.7) is sufficient. □

REFERENCES

[1] S. S. Bonan, *Weighted mean convergence of Lagrange interpolation*, Ph.D. thesis, Ohio State University, Columbus, OH, 1982.

[2] Z. Ditzian and V. Totik, *Moduli of Smoothness*, *Springer Series in Computational Mathematics*, vol. 9, Springer-Verlag, Berlin, 1987.

[3] G. Freud, *Orthogonal Polynomials*, Pergamon Press/Akademiai Kiado, Oxford/Budapest, 1971.

[4] A. Knopfmacher and D. S. Lubinsky, *Mean convergence of Lagrange interpolation for Freud's weights with application to product integration rules*, J. Comp. Appl. Math., 17 (1987), pp. 79–103.

[5] A. L. Levin and D. S. Lubinsky, *Christoffel Functions, Orthogonal Polynomials, and Nevai's Conjecture for Freud Weights*, Constr. Approx., 8 (1992), pp. 461–533.

[6] D. S. Lubinsky, A. Mate, and P. Nevai, *Quadrature sums involving pth powers of polynomials*, SIAM J. Math. Anal., 18 (1987), pp. 531–544.

[7] D. S. Lubinsky and F. Moricz, *The Weighted L_p-Norms of Orthonormal Polynomials for Freud Weights*, J. Approx. Theory, 77 (1994), pp. 42–50.

[8] D. S. Lubinsky and A. Sidi, *Convergence of product integration rules for functions with interior and endpoint singularities over bounded and unbounded intervals*, Math. Comp., 46 (1986), pp. 297–313.

[9] H. N. Mhaskar and E. B. Saff, *Extremal problems for polynomials with exponential weights*, Trans. Amer. Math. Soc., 285 (1984), pp. 203–234.

[10] ———, *Where does the sup-norm of a weighted polynomial live?*, Constr. Approx., 1 (1985), pp. 71–91.

[11] B. Muckenhoupt, *Mean convergence of Hermite and Laguerre series* II, Trans. Amer. Math. Soc., 147 (1970), pp. 433–460.

[12] P. Nevai, *Orthogonal Polynomials*, Mem. Amer. Math. Soc., 213 (1979).

[13] ———, *Mean convergence of Lagrange interpolation*, II, J. Approx. Theory, 30 (1980), pp. 263–276.

[14] ———, *Geza Freud, orthogonal polynomials and Christoffel functions: A case study*, J. Approx. Theory, 48 (1986), pp. 3–167.

[15] J. Szabados and P. Vertesi, *Interpolation of Functions*, World Scientific, Singapore, 1991.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 263–279, March 1995

INVERSE PROBLEMS AT THE BOUNDARY FOR AN ELASTIC MEDIUM *

GEN NAKAMURA† AND GUNTHER UHLMANN‡

Abstract. In this paper, it is proven that one can determine the full Taylor series of the elastic tensor of an elastic, isotropic, inhomogeneous medium in all dimensions $n \geq 2$ and for a generic anisotropic elastic tensor in two dimensions by making measurements at the boundary of the medium of the displacement vectors and corresponding stresses. This information is encoded in the so-called Dirichlet-to-Neumann map.

Key words. inverse boundary problems, elasticity tensor, Dirichlet-to-Neumann map

AMS subject classifications. 35R30, 35P05, 35J05

1. Introduction and statement of the results.

Suppose we have a linear, inhomogeneous elastic medium. The inverse problem we address in this paper is whether knowledge of the displacement and the corresponding stress at the boundary of the medium determines its elastic parameters. In [N-U], the authors proved that the answer is in the affirmative in two dimensions in the case that the medium is isotropic and the Lamé parameters are close to constant in an appropiate topology. In this paper, we prove that from the boundary information given one can determine the Taylor series of the Lamé parameters at the boundary in the isotropic case in all dimensions. We also prove a similar result for anisotropic conductivities in two dimensions which satisfy additional conditions. The boundary determination of the elastic parameters in the isotropic case in two dimensions has been proved earlier in [A-N-S]. In a forthcoming article, we prove that one can determine the Lamé parameters in the interior by making the same type of boundary measurements in all dimensions $n \geq 3$ (see [N-UI]). The result proven in this paper is a necessary ingredient in the proof of the identifiability result in the interior.

Now we state more precisely the problem and the main results. Let $n \geq 2$ be an integer and $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary $\partial\Omega$. Physically, Ω is considered as a linear, inhomogeneous, elastic medium. The deformation of Ω due to the displacement vector $\vec{f} \in C^\infty(\partial\Omega)$ given on $\partial\Omega$ is expressed by the following Dirichlet boundary value problem in terms of the displacement vector $\vec{u} = \vec{u}(x)$:

$$\begin{cases} L_C \vec{u} &= \operatorname{div} \sigma(\vec{u}) = 0 \text{ in } \Omega, \\ \vec{u}\,|_{\partial\Omega} &= \vec{f}, \end{cases} \tag{1.1}$$

where $\sigma(\vec{u})$ is the stress tensor given by the generalized Hooke's law:

$$\sigma\left(\vec{u}\right) = C\varepsilon\left(\vec{u}\right). \tag{1.2}$$

*Received by the editors April 14, 1993; accepted for publication (in revised form) September 28, 1993.

†Department of Mathematics, Science University of Tokyo, Shinjuku-ku, Tokyo 162, Japan.

‡Department of Mathematics, University of Washington, Seattle, Washington, 98195. This research was partially supported by National Science Foundation grant DMS 91-00178.

The componentwise expression of (1.2) is

$$\sigma_{ij}(\vec{u}) = \sum_{k,\ell=1}^{n} C_{ijk\ell}\varepsilon_{k\ell}(\vec{u}) \quad (1 \le i, j \le n). \tag{1.3}$$

Here $\varepsilon(\vec{u})$ is the linear strain given by

$$\varepsilon(\vec{u}) := \operatorname{Sym}\nabla \vec{u} = 2^{-1}(\nabla \vec{u} + (\nabla \vec{u})^T), \tag{1.4}$$

where $C = C(x) = \left(C_{ijk\ell}(x)\right)_{1\le i,j,k,\ell\le n} \in C^\infty(\overline{\Omega})$ denotes the elastic tensor and $(\nabla \vec{u})^T$ denotes the transpose of $\nabla \vec{u}$.

We shall assume throughout the paper that the elastic tensor satisfies the following symmetries and strong convexity, which are physically natural conditions:

Symmetries. For any i, j, k, ℓ $(1 \le i, j, k, \ell \le n)$,

$$\text{(i) } C_{ijk\ell} = C_{ij\ell k}, \quad \text{(ii) } C_{ijk\ell} = C_{jik\ell}, \quad \text{(iii) } C_{ijk\ell} = C_{k\ell ij}. \tag{1.5}$$

Strong convexity. There exists $\delta > 0$ such that

$$\operatorname{trace}(\varepsilon(C\varepsilon)) \ge \delta\|\varepsilon\|^2 \tag{1.6}$$

on $\overline{\Omega}$ for any $n \times n$ matrix ε.

It is well known that (1.1) admits a unique solution $\vec{u} \in C^\infty(\overline{\Omega})$ if C satisfies (1.5) and (1.6). We define $\Lambda_C(\vec{f}) \in C^\infty(\partial\Omega)$ by

$$\Lambda_C(\vec{f}) = \sigma(\vec{u})\,\vec{\nu}\,|_{\partial\Omega}, \tag{1.7}$$

where $\vec{u}$ is the solution of (1.1) and $\vec{\nu}$ denotes the unit outer normal field to $\partial\Omega$.

DEFINITION 1.8. *We call the map*

$$\Lambda_C : C^\infty(\partial\Omega) \ni \vec{f} \longmapsto \Lambda_C(\vec{f}) \in C^\infty(\partial\Omega)$$

the Dirichlet-to-Neumann map (D-N).

The problem we consider in this paper is whether we can recover C and all of its derivatives on $\partial\Omega$ from Λ_C. In this paper, we prove that this is the case for isotropic n-dimensional medium and for a "generic" class of anisotropic elastic tensors in two dimensions.

DEFINITION 1.9. *We call the elastic medium Ω or its elastic tensor C isotropic if $C = (C_{ijk\ell})_{1\le i,j,k,\ell\le n}$ is given by*

$$C_{ijk\ell} = \lambda\delta_{ij}\delta_{k\ell} + \mu(\delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk})$$

with Lamé moduli λ, $\mu \in C^\infty(\overline{\Omega})$, where δ_{ik} is the Kronecker delta.

THEOREM 1.10. *For any $n \ge 2$, for any isotropic elastic medium Ω there is an inversion formula for identifying C and all of its derivatives on $\partial\Omega$ from Λ_C.*

Remark 1.11.

(i) The inversion formula can be seen in the proof of Theorem 1.10, which is given in §3.

(ii) Akamatsu, Nakamura, and Steinberg [A-N-S] proved Theorem 1.10 for $n = 2$.

We now describe a corresponding result for anisotropic C in two dimensions under some conditions that we formulate below.

Let $(x^1, \dots, x^n)$ denote local coordinates and $(x_1, \dots, x_n)$ denote cartesian coordinates. Set

$$C^{ijk\ell}(x) = \sum_{a,b,c,d=1}^{n} g^{ai}(x)g^{bj}(x)g^{ck}(x)g^{d\ell}(x)C_{abcd}(x)$$

$$\varepsilon^{ij}(x) = \sum_{a,b=1}^{n} g^{ai}(x)g^{bj}(x)\varepsilon_{ab}(x),$$

where

$$g^{ai}(x) = \sum_{r=1}^{n} \frac{\partial x^a}{\partial x_r}(x)\frac{\partial x^i}{\partial x_r}(x).$$

Assume that there exist scalar functions $\lambda_\alpha \in C^\infty(\overline{\Omega})$ $(1 \le \alpha \le 2n)$ satisfying

$$\sum_{\alpha,\beta=1}^{n(n+1)/2} C^{\alpha,\beta}\varepsilon_\alpha\varepsilon_\beta = \sum_{\alpha=1}^{n(n+1)/2} \lambda_\alpha\varepsilon_\alpha\varepsilon^\alpha \tag{1.12}$$

for any covariant symmetric tensor (ε_α) and the associated contravariant symmetric tensor (ε^α). Here $(C^{\alpha,\beta})$, (ε_α), (ε^α) are defined by renumerating the double indices (i, j), (k, ℓ) of $(C^{ijk\ell})$, (ε_{ij}), (ε^{ij}) into single indices α, β.

THEOREM 1.13. *Let $n = 2$ and C satisfy* (1.12) *and the generic condition in Definition* (2.5). *Then there is an inversion formula for identifying C and all of its derivatives on $\partial\Omega$ from Λ_C.*

Remark 1.14. The inversion formula can be seen in the proof of Theorem 1.13 in §3.

The proof of the main results follows from the asymptotic formula for the full symbol of the D-N map proven in §2. We write the full symbol of the D-N map in terms of the surface impedance tensor. In the isotropic case we can easily recover the elastic tensor from the surface impedance tensor (see §3). For the anisotropic case in two dimensions the surface impedance tensor has three components and the elastic tensor has also three components under the assumption (1.12), so that it becomes possible to determine one from the other. This is the main reason to assume a condition like (1.12). The computations in §2 rely on Stroh's formalism which can be seen, for instance, in [C-S] for the three-dimensional case. We include in the appendix an extension of Stroh's formalism to the n-dimensional case.

We also note that the problem considered here is a direct analog of the inverse conductivity problem. Theorem 1.10 is the analog of a corresponding result proven by Kohn and Vogelius [K-V] for the inverse conductivity problem. Another proof of this result using the full symbol of ther D-N map was given in [S-U]. The boundary determination in the anisotropic case was proven in [L-U].

2. Full symbol of the D-N map. In what follows we identify the orthogonal complement $N^*(\partial\Omega)^\perp$ of the conormal bundle $N^*(\partial\Omega)$ of $\partial\Omega$ and the cotangent bundle $T^*(\partial\Omega)$ of $\partial\Omega$ by giving the Euclidean metric to $\mathbb{R}^n$ and its induced metric to $\partial\Omega$. The identification is done via the unitary map

$$N_y^*(\partial\Omega)^\perp \ni \xi \longmapsto \xi|_{T_y^*(\partial\Omega)} \in T_y^*(\partial\Omega)$$

for each $y \in \partial\Omega$. Moreover, we assume that the elasticity tensor C satisfies (1.5) and (1.6).

Let $s(x)$ be a local defining function of $\partial\Omega$ such that $|ds(x)| = 1$ on $\partial\Omega$ and $x = (x^1, \dots, x^{n-1}, x^n) = (y^1, \dots, y^{n-1}, s) = y$ be local coordinates such that $dy^j \perp ds (1 \leq j \leq n-1)$ near a fixed point $y_0 \in \partial\Omega$. We assume that Ω is locally given by $\{s(x) > 0\}$.

DEFINITION 2.1. *For any $(y, \eta) \in T^*(\partial\Omega)$, we define the surface impedance tensor $Z(y, \eta)$ by*

$$\begin{cases} Z(y,\eta) &= -Q^{-1}(y,\eta) - \sqrt{-1}Q^{-1}(y,\eta)S(y,\eta), \\ Q(y,\eta) &= -(2\pi)^{-1}\int_0^{2\pi} \langle w, w\rangle^{-1} d\varphi, \\ S(y,\eta) &= -(2\pi)^{-1}\int_0^{2\pi} \langle w, w\rangle^{-1}\langle w, \zeta\rangle d\varphi. \end{cases} \tag{2.2}$$

where the (j, k) component $\langle \zeta, w\rangle_j^k$ of the mixed tensor $\langle \zeta, w\rangle$ is given by

$$\langle \zeta, w\rangle_j^k = \sum_{i,\ell=1}^{n} \zeta_i C_j^{i\ell k}(y) w_\ell, \quad C_j^{i\ell k}(y) = \sum_{a,b,c=1}^{n} g^{ai}(y) g^{b\ell}(y) g^{ck}(y) C_{ajbc}(y),$$

where

$$\zeta = (\cos\varphi)|\eta|^{-1}\eta + (\sin\varphi) ds(y) = \sum_{i=1}^{n} \zeta_i (dx^i)_y,$$

$$w = -(\sin\varphi)|\eta|^{-1}\eta + (\cos\varphi) ds(y) = \sum_{i=1}^{n} w_i (dx^i)_y.$$

Similarly, we can define the surface impedance tensor $Z(\overline{y}, \overline{\eta})$ for every surface $\Gamma(\overline{y}) : s(x; \overline{y}) := s(x) - s(\overline{y}) = 0$, where $\overline{y}$ is close to $y \in \partial\Omega$ and $(\overline{y}, \overline{\eta}) \in T^*_{\overline{y}}(\Gamma(\overline{y}))$. By the natural identification $T^*_y(\partial\Omega) \cong T^*_{\overline{y}}(\Gamma(\overline{y})) \cong \mathbb{R}^n$, we can fix $\eta \in T^*_y(\partial\Omega) \cong T^*_{\overline{y}}(\Gamma(\overline{y}))$ and let $Z(\overline{y}, \eta)$ depend smoothly on $\overline{y}$ near $\overline{y} = y$. If any variation along the normal direction of $\partial\Omega$ is necessary, we interpret the surface impedance tensor in this manner. For example, in Theorem 2.6 we consider the derivatives of $Z(y, \eta)$ with respect to s.

DEFINITION 2.3. For any $(y, \eta) \in T^*(\partial\Omega)\backslash 0$, we define $K_s(y, \eta) \in C^\infty([0, \varepsilon); \mathrm{Hom}(\Xi, \Xi))$[1] by

$$K_s(y,\eta) = \begin{bmatrix} K_s^{11}(y,\eta), & K_s^{12}(y,\eta) \\ K_s^{21}(y,\eta), & K_s^{22}(y,\eta) \end{bmatrix} \quad (s \in [0, \varepsilon)), \tag{2.4}$$

with

$$\begin{cases} K_s^{11}(y,\eta) &= -\langle ds(x), ds(x)\rangle^{-1}\langle ds(x), \eta\rangle \\ K_s^{12}(y,\eta) &= -|\eta|\langle ds(x), ds(x)\rangle^{-1} \\ K_s^{21}(y,\eta) &= -|\eta|^{-1}\{\langle \eta, ds(x)\rangle\langle ds(x), ds(x)\rangle^{-1}\langle ds(x), \eta\rangle - \langle \eta, \eta\rangle\} \\ K_s^{22}(y,\eta) &= -\langle \eta, ds(x)\rangle\langle ds(x), ds(x)\rangle^{-1} \end{cases}$$

[1]From now on, $C^\infty(U; V)$ (resp., $C^\infty(U, V)$) denotes the set of smooth V-valued functions (resp., sections $U \longrightarrow V$).

and $\Xi = \pi^*(T^*(\partial\Omega)^{\mathbb{C}} \oplus \pi^*(T^*(\partial\Omega))^{\mathbb{C}}$ with π the natural projection $\pi : T^*(\partial\Omega) \longrightarrow \partial\Omega$. Here "$\oplus$," "$\mathbb{C}$," and "Hom" denote the Whitney sum, complexification, and the vector bundle of homomorphisms, respectively. Moreover, $x = (y, s)$ and $\langle ds(x), ds(x)\rangle$, $\langle ds(x), \eta\rangle$ are defined similarly as the above $\langle \zeta, w\rangle$ by using the identification $\eta \in T_y(\partial\Omega) \cong T_x(\Gamma(x))$.

DEFINITION 2.5. *We say that the elasticity tensor C is* generic *at $(y_0, \eta^0) \in T^*(\partial\Omega)\backslash 0$ if all the eigenvalues of $K_s(y, \eta)$ are distinct for each $s \in [0, \varepsilon)$ and (y, η) in a conic neighborhood (y_0, η^0).*

It is well known that the D-N map

$$\Lambda_C : C^\infty(\partial\Omega, T^*(\partial\Omega)^{\mathbb{C}} \oplus T^*(\partial\Omega)^{\mathbb{C}}) \longrightarrow C^\infty(\partial\Omega, T^*(\partial\Omega)^{\mathbb{C}} \oplus T^*(\partial\Omega)^{\mathbb{C}})$$

is a classical pseudodifferential operator of order 1.

In terms of the above local coordinates $(x^1, \ldots, x^{n-1}, x^n) = (y^1, \ldots, y^{n-1}, s)$, we have the following asymptotic representation formula for the full symbol $\tilde{\sigma}(\Lambda_C)$ of Λ_C.

THEOREM 2.6. *Assume that either* (i) *or* (ii) *is satisfied.*

(i) *C is generic at (y_0, η^0).*

(ii) *C is isotropic near y_0.*

Then we have the following formula for $\tilde{\sigma}(\Lambda_C)(y, \eta)$:

$$\tilde{\sigma}(\Lambda_C)(y, \eta) \simeq -|\eta| Z(y, \eta) + \sum_{j=1}^{\infty} |\eta|^{1-j} W_j (D_s^j Z)(y, \eta) \tag{2.7}$$

in a conic neighborhood of (y_0, η^0), where $D_s = -\sqrt{-1}\frac{\partial}{\partial s}$ and each W_j is a linear bijective map on the set of all $n \times n$ matrices which do not depend on the s derivatives of C. The meaning of "$\simeq$" in (2.7) is as follows. For each $k \in \mathbb{N}$,

$$\tilde{\sigma}(\Lambda_C)(y, \eta) + |\eta| Z(y, \eta) - \sum_{j=1}^{k-1} |\eta|^{1-j} W_j (D_s^j Z)(y, \eta) = 0 \mod (\tilde{T}^{k-1}, S^{-k+1})$$

in a conic neighborhood of (y_0, η^0). Here $\mathrm{mod}(\tilde{T}^{k-1}, S^{-k+1})$ means that we are neglecting the terms in S^{-k+1} in a conic neighborhood (y_0, η^0) and the term $\notin S^{-k+1}$ in a conic neighborhood which depend only on the s derivatives of C up to order $k-1$. Moreover, S^{-k+1} is used to denote either the usual class of classical symbols or the associated class of pseudodifferential operators and $\sum_{k=1}^{n-1}$ does not exist for $n = 1$.

Proof. We first give an outline of the computation of $\tilde{\sigma}(\Lambda_C)(y, \eta)$.

Step 1. We write (1.1) locally in the following form using tensorial notation.

$$\begin{cases} M\vec{u} = 0 & \text{in } \Omega, \\ \vec{u}\,|_{\partial\Omega} = \vec{f}, \end{cases} \tag{2.8}$$

where

$$M\cdot := \left(\sum_{i,\ell=1}^{n} \nabla_i \left(C_j^{i\ell k}(x) \nabla_\ell \cdot \right) \quad \substack{j \downarrow 1,\ldots,n \\ k \rightarrow 1,\ldots,n} \right), \tag{2.9}$$

$$B\cdot := \left(\sum_{i,\ell=1}^{n} C_j^{i\ell k}(x) \frac{\partial s}{\partial x^i}(x) \nabla_\ell \cdot \quad \substack{j \downarrow 1,\ldots,n \\ k \rightarrow 1,\ldots,n} \right), \tag{2.10}$$

and ∇_i is the covariant differential with respect to $\frac{\partial}{\partial x^i}$.

Step 2. We transform (2.8) into the following boundary value problem:

$$\begin{cases} D_s U = N_s U & \text{in } s > 0, \\ [I_n, 0_n] U|_{s=0} & = \Lambda \vec{f}, \end{cases} \tag{2.11}$$

where

$$U = \begin{bmatrix} \Lambda \vec{u} \\ \vec{v} \end{bmatrix}, \qquad \tilde{\sigma}(\Lambda)(y,\eta) = |\eta|, \tag{2.12}$$

$$\vec{v} = -\langle ds, ds\rangle D_s \vec{u} - \langle ds, D_y\rangle \vec{u}, \quad \tilde{\sigma}(\langle ds, D_y\rangle) = \langle ds, \eta\rangle, \tag{2.13}$$

$$K_s(y,\eta) = \sigma(N_s)(y,\eta) \text{ (i.e., the principal symbol of } N) \tag{2.14}$$

and $[I_n, 0_n]$ denotes the matrix whose first n rows and columns is the identity matrix and whose last n rows and n columns is the zero matrix.

Step 3. We diagonalize the system. There exists classical pseudodifferential operators $Q_s(y, D_y)$ and $\widetilde{N}_s(y, D_y)$ of order 0 and 1, respectively, depending smoothly on $s \in [0, \varepsilon)$, such that

$$LQ_s - Q_s \widetilde{L} = 0 \mod S^{-\infty}, \tag{2.15}$$

where

$$L = D_s - N_s, \quad \widetilde{L} = D_s - \widetilde{N}_s, \quad \tilde{\sigma}(\tilde{N}_s) = \begin{bmatrix} \tilde{N}_s^+, & 0_n \\ 0_n, & \tilde{N}_s^- \end{bmatrix} \tag{2.16}$$

with $\pm$ (the imaginary part of each eigenvalue of $\sigma(\tilde{N}_s^{\pm})) > 0$, and

$$\tilde{\sigma}(Q_s) \sim \sigma(Q_s) + \sum_{j=1}^{\infty} R_s^{(-j)}(y,\eta), \quad \text{ord } R_s^{(-j)} = -j \tag{2.17}$$

and the order of the s derivatives of C in each $R_s^{(-j)}$ is not greater than j.

Consider the well-posed Cauchy problem:

$$\begin{cases} (D_s - \widetilde{N}_s^+) \vec{w}^+ = 0 & \text{in } s > 0, \\ \vec{w}^+|_{s=0} = \vec{h}. \end{cases} \tag{2.18}$$

We define the map T by

$$T\vec{h} := \vec{w}^+. \tag{2.19}$$

Step 4. We write the D-N map in terms of Q_s. Clearly,

$$U := Q_s \tilde{J} T \vec{h}, \quad \tilde{J} = \begin{bmatrix} I_n \\ 0_n \end{bmatrix}$$

satisfies

$$(D_s - N_s)U = 0 \bmod C^\infty.$$

Moreover, if $\vec{h}$ satisfies

$$R\vec{h} = \Lambda \vec{f}, \qquad R = ([I_n, 0_n]Q_s\tilde{J}T)|_{s=0}, \tag{2.20}$$

then U satisfies the boundary condition of (2.11). Hence the Poisson operator P is given by

$$P = Q_s\tilde{J}TR^{-1}\Lambda \mod S^{-\infty}. \tag{2.21}$$

Therefore, the D-N map Λ_C is given by

$$\Lambda_C = -\sqrt{-1}[0_n, I_n]P|_{s=0} \mod S^{-\infty}. \tag{2.22}$$

Although (2.22) is a global formula for Λ_C, we only have to compute $\tilde{\sigma}(\Lambda_C)$ in a conic neighborhood of (y_0, η^0) (i.e., microlocally at (y_0, η^0)). Hence in the following we assume that all pseudodifferential operators are defined microlocally at (y_0, η^0) and the the notation for pseudodifferential operators and their symbols such as "$\cdot \in S^{-k+1}$," "$\cdot \equiv \cdot \bmod S^{-k}$," "$\cdot \sim \cdot$," "$\cdot \simeq \cdot$," "$\cdot = \cdot$," "$\bmod(\tilde{T}^{k-1}, S^{-k+1})$," etc. are understood to be microlocal at (y_0, η^0).

We now go into detail of the proof of Theorem 2.6. To start with, we explain more precisely the choice of Q_s and $\tilde{N}_s$.

For simplicity we drop the subscript "s" from now on. So for example, we will simply write Q instead of Q_s.

Since L inherits the ellipticity from that of M, there exists a homogeneous elliptic symbol $\tilde{Q}(y, \eta)$ of order 0 such that

$$K(y, \eta)\tilde{Q}(y, \eta) = \tilde{Q}(y, \eta)\sigma(\tilde{N})(y, \eta). \tag{2.23}$$

Hence there exists $D \in S^0$ such that

$$L\tilde{Q} - \tilde{Q}L' \equiv 0 \bmod S^0, \tag{2.24}$$

where

$$L' = D_s - \Lambda - D, \qquad \Lambda = \sigma(\tilde{N})(y, D_y), \tag{2.25}$$

$$\sigma(D) \equiv -\tilde{Q}^{-1}D_s\tilde{Q} \mod (\tilde{T}^0, S^{-1}). \tag{2.26}$$

Next we seek a pseudodifferential operator $\Phi \in S^0$ such that

$$\Phi - I_{2n} \in S^{-1}, \qquad L'\Phi - \Phi\tilde{L} \equiv 0 \bmod S^{-\infty}. \tag{2.27}$$

Hence Q can be chosen as

$$Q = \tilde{Q}\Phi \tag{2.28}$$

and $\tilde{N}$ is constructed in the process of choosing Q.

In order to construct Q and $\widetilde{N}$ we define

$$\Phi^{(-j)} = \begin{bmatrix} 0_n & \Phi_{12}^{(-j)} \\ \Phi_{21}^{(-j)} & 0_n \end{bmatrix} \in S^{-j} \quad (j = 1, 2, \ldots) \tag{2.29}$$

and

$$E^{(-j)} = \begin{bmatrix} E_+^{-(j)} & 0_n \\ 0_n & E_-^{(-j)} \end{bmatrix} \in S^{-j} \quad (j = 0, 1, 2, \ldots) \tag{2.30}$$

inductively as follows. Set

$$J^{(0)} = L' I_{2n} - I_{2n}(D_s - \Lambda) = \begin{bmatrix} J_{11}^{(0)} & J_{12}^{(0)} \\ J_{21}^{(0)} & J_{22}^{(0)} \end{bmatrix}. \tag{2.31}$$

Define $\Phi^{(-1)}$ and $E^{(0)}$ by

$$\begin{cases} \widetilde{N}^+ \Phi_{12}^{(-1)} - \Phi_{12}^{(-1)} \widetilde{N}^- - \sigma(J_{12}^{(0)}) = 0, \\ \widetilde{N} \Phi_{21}^{(-1)} - \Phi_{21}^{(-1)} \widetilde{N}^+ - \sigma(J_{21}^{(0)}) = 0, \end{cases} \tag{2.32}$$

and

$$E^{(0)} = -\begin{bmatrix} \sigma(J_{11}^{(0)}) & 0_n \\ 0_n & \sigma(J_{22}^{(0)}) \end{bmatrix}. \tag{2.33}$$

Having constructed $\Phi^{(-j)} (1 \le j \le k)$, $E^{(-j)} (0 \le j \le k)$, we define $\Phi^{(-k-1)}$, $E^{(-k)}$ by

$$\begin{cases} \widetilde{N}^+ \Phi_{12}^{(-k-1)} - \Phi_{12}^{(-k-1)} \widetilde{N}^- - \sigma(J_{12}^{(-k)}) = 0, \\ \widetilde{N}^- \Phi_{21}^{(-k-1)} - \Phi_{21}^{(-k-1)} \widetilde{N}^+ - \sigma(J_{21}^{(-k)}) = 0, \end{cases} \tag{2.34}$$

$$E^{(-k)} = -\begin{bmatrix} \sigma(J_{11}^{(-k)}) & 0_n \\ 0_n & \sigma(J_{22}^{(-k)}) \end{bmatrix}, \tag{2.35}$$

where

$$\begin{aligned} J^{(-k)} &= \begin{bmatrix} J_{11}^{(-k)} & J_{12}^{(-k)} \\ J_{21}^{(-k)} & J_{22}^{(-k)} \end{bmatrix} \\ &= L'\Big(I + \sum_{j=1}^{k} \Phi^{(-j)}\Big) - \Big(I + \sum_{j=1}^{k} \Phi^{(-j)}\Big)\Big(D_s - \Lambda - \sum_{j=0}^{k-1} E^{(-j)}\Big). \end{aligned} \tag{2.36}$$

Here we note that (2.32) and (2.34) are uniquely solvable for $\Phi_{12}^{(-1)}$, $\Phi_{21}^{(-1)}$, and $\Phi_{12}^{(-k-1)}$, $\Phi_{21}^{(-k-1)}$, because $\widetilde{N}^+$ and $\widetilde{N}^-$ do not have any common eigenvalues.

We prove now the first key lemma.

LEMMA 2.37.

$$\begin{cases} S^0 \ni J^{(0)} \equiv \widetilde{Q}^{-1} D_s \widetilde{T} \mod (\widetilde{T}^0, S^{-1}), \\ S^{-k} \ni J^{(-k)} \equiv D_s \Phi^{(-k)} \mod (\widetilde{T}^k, S^{-k-1}) \quad (k = 1, 2, \ldots), \end{cases} \tag{2.38}$$

and

$$\text{(2.39)}\qquad \begin{cases} \Phi^{(-k-1)}(y,\eta) & \in \widetilde{T}^{k+1}\cdot S^{-k-1}, \\ E^{(-k)}(y,\eta) & \in \widetilde{T}^{k+1}\cdot S^{-k} \end{cases}$$

for any $k = 0,1,2,\ldots$. *Here, for any* $0 \le \ell \in \mathbb{Z}$, $m \in \mathbb{Z}$, $\widetilde{T}^{\ell}\cdot S^{m}$ *denotes the homogeneous classical symbols of order* m *which depend only on the* s *derivatives of* C *up to order* ℓ.

Proof. It is easy to see that

$$\begin{cases} S^0 \ni J^{(0)} \equiv \widetilde{Q}^{-1}D_s\widetilde{Q} \bmod (\widetilde{T}^0,\, S^{(-1)}), \\ \Phi^{(-1)}(y,\eta) \in \widetilde{T}^1\cdot S^{-1},\ E^{(0)}(y,\eta) \in \widetilde{T}^1\cdot S^0. \end{cases}$$

Then

$$\text{(2.40)}\qquad \begin{cases} S^{-k} \ni J^{(-k)} \equiv D_s\Phi^{(-k)} \bmod (\widetilde{T}^k, S^{-k-1}), \\ \Phi^{(-k-1)}(y,\eta) \in \widetilde{T}^{k+1}\cdot S^{-k-1}, \qquad E^{(-k)}(y,\eta) \in \widetilde{T}^{k+1}\cdot S^{-k} \end{cases}$$

for any $k = 1,2,\ldots$ is proved by induction on k.

Since it is easy to prove that (2.40) is already valid for $k = 1$, we only need to prove (2.40) for $k+1$, assuming that it is already valid for k. Observe that

$$\begin{aligned} &J^{(-k-1)} \\ &= L'\Big(I_{2n} + \sum_{j=1}^{k+1}\Phi^{(-j)}\Big) - \Big(I_{2n} + \sum_{j=1}^{k+1}\Phi^{(-j)}\Big)\Big(D_s - \Lambda - \sum_{j=0}^{k}E^{(-j)}\Big) \\ &= \Big\{L'\Big(I_{2n} + \sum_{j=1}^{k}\Phi^{(-j)}\Big) - \Big(I_{2n} + \sum_{j=1}^{k}\Phi^{(-j)}\Big)\Big(D_s - \Lambda - \sum_{j=0}^{k-1}E^{(-j)}\Big)\Big\} + L'\Phi^{(-k-1)} \\ &\quad - \Phi^{(-k-1)}\Big(D_s - \Lambda - \sum_{j=0}^{k}\Phi^{(-j)}\Big) + \Big(I_{2n} + \sum_{j=1}^{k+1}\Phi^{(-j)}\Big)E^{(-k)} \\ &= \Big\{J^{(-k)} - (\Lambda\Phi^{(-k-1)} - \Phi^{(-k-1)}\Lambda)\Big\} + D_s\Phi^{(-k-1)} + \Big\{\Phi^{(-k-1)}\sum_{j=0}^{k}E^{(-j)} \\ &\qquad\qquad + \sum_{j=1}^{k+1}\Phi^{(-j)}E^{(-k)}\Big\}. \end{aligned}$$

Hence

$$S^{-k-1} \ni J^{-k-1} \equiv D_s\Phi^{(-k-1)} \bmod (\widetilde{T}^{k+1},\, S^{-k-2}).$$

The facts that $\Phi^{(-k-2)}(y,\eta) \in \widetilde{T}^{k+2}\cdot S^{-k-2}$ and $E^{(-k-1)}(y,\eta) \in \widetilde{T}^{k+2}\cdot S^{-k-1}$ are clear from (2.34) and (2.35).

Now let $\widetilde{N}_s$ and Φ_s be pseudodifferential operators of order 1 and 0, respectively, depending smoothly on $s \in [0,\varepsilon)$ such that

$$\text{(2.41)}\qquad \widetilde{\sigma}(\widetilde{N}_s) \sim \Lambda + \sum_{j=0}^{\infty}E_s^{(-j)}, \qquad \Phi_s \sim I_{2n} + \sum_{j=1}^{\infty}\Phi_s^{(-j)}.$$

Then, using (2.36) and (2.38), it is easy to see that Q_s and $\widetilde{N}_s$ satisfy (2.15)–(2.17).

Now we arrange our local coordinates $(x^1, \dots, x^{n-1}, x^n)$ so that

$$\eta^0/|\eta^0| = (1, 0, \dots, 0), \qquad ds(y_0) = (0, \dots, 0, 1) \tag{2.42}$$

and $dx^j (1 \le j \le n)$ are orthonormal at y_0 in terms of these special local coordinates. Then, as in [N], $K(y_0, \eta^0)$ is related to $N(0)$, which is defined by (4.45) in the Appendix through the relation

$$K(y_0, \eta^0) = \begin{bmatrix} -I_n, & 0_n \\ 0_n, & I_n \end{bmatrix} N(0) \begin{bmatrix} -I_n & 0_n \\ 0_n & I_n \end{bmatrix} \tag{2.43}$$

in terms of these coordinates. By using (2.43), we normalize the eigenvectors and the generalized eigenvectors $\langle -a_\alpha, \ell_\alpha \rangle$ $(1 \le \alpha \le 2n)$ of $K(y_0, \eta^0)$ in terms of these coordinates using (4.8) in [C-S]. Here a_α, ℓ_α are similar to those in (4.14) of [C-S].

Let $A_\alpha, L_\alpha (1 \le \alpha \le 2n)$ be covariant vectors such that

$$A_\alpha = a_\alpha, \quad L_\alpha = \ell_\alpha \tag{2.44}$$

in terms of the special local coordinates, and define the $n \times n$ matrices A, L by

$$A = [A_1, \dots, A_n], \qquad L = [L_1, \dots, L_n]. \tag{2.45}$$

Then we can define $\widetilde{Q}$ by

$$\widetilde{Q} = \begin{bmatrix} -A, & -\overline{A} \\ L, & \overline{L} \end{bmatrix}, \tag{2.46}$$

where "$-$" denotes complex conjugate.

Let A^α, L^α $(1 \le \alpha \le 2n)$ be the contravariant vectors associated with A_α, L_α $(1 \le \alpha \le 2n)$ and define the $n \times n$ matrices A', L' by

$$A' = [A^1, \dots, A^n], \qquad L' = [L^1, \dots, L^n]. \tag{2.47}$$

If we interpret (4.8) in [C-S] by using the tranformation rule of covariant and contravariant vectors under change of local coordinates, we have

$$\widetilde{Q}^T J \widetilde{Q}' = I, \tag{2.48}$$

where

$$\widetilde{Q}' = \begin{bmatrix} -A' & -\overline{A}' \\ L' & \overline{L}' \end{bmatrix}, \qquad J = \begin{bmatrix} 0_n & I_n \\ I_n & 0_n \end{bmatrix}. \tag{2.49}$$

Hence

$$\widetilde{Q}^{-1} = (\widetilde{Q}')^T J. \tag{2.50}$$

It is convenient now to introduce the following definition.

DEFINITION 2.51.

(i) *For a given matrix* $F = (F_{jk})_{1\le j,k\le n}$, *we denote the unique solution* X *of the equation* $\widetilde{N}^+X - X\widetilde{N}^- = F$ *by* $X = W(F)$.

(ii) *For* $a, b \in \mathbb{Z}$, $a \ge 0$, *we define* $\sum_{\ell=0}^{\infty} \widetilde{T}^{\ell+a}.S^{b-\ell} = \{$*classical symbols* $r_s(y,\eta)$ *which depend smoothly on* $s \in [0,\varepsilon)$ *and admit an asymptotic expansion* $r_s(y,\eta) \sim \sum_{\ell=0}^{\infty} r_s^{b-\ell}(y,\eta)$ *such that* $r_s(y,\eta) \simeq \sum_{\ell=0}^{\infty} r_s^{b-\ell}(y,\eta)\}$.

Here " $\simeq$ *" means that each* $r_s^{b-\ell}(y,\eta)$ *depends only on the* s *derivatives of* C *up to order* $\ell+a$ *and* $r_s - \sum_{\ell=0}^{k-1} r^{b-\ell} \equiv 0 \bmod (\widetilde{T}^{k+a-1}, S^{b-k})$ *for each* $k \in \mathbb{N}$. *In particular,* $\widetilde{T}^{\ell+a} \cdot S^{b-\ell} = \{$*homogeneous classical symbols* $r_s^{b-\ell}(y,\eta)$ *of order* $b-\ell$ *which depend smoothly on* $s \in [0,\varepsilon)$ *and only on the* s *derivatives of* C *up to order* $\ell+a\}$.

The following is also a key lemma.

LEMMA 2.52. *For each* $j = 1, 2, \ldots$, *we have*

$$\Phi_{12}^{(-j)} \equiv -W^j\big((L')^T D_s^j \overline{A} + (A')^T D_s^j \overline{L}\big) \mod \big(\widetilde{T}^{j-1}, S^{-j-1}\big) \tag{2.53}$$

and

$$\overline{\Phi_{21}^{(-j)}} \equiv (-1)^j \Phi_{12}^{(-j)} \mod (\widetilde{T}^{j-1}, S^{-j-1}). \tag{2.54}$$

Proof. Both statements are proved by induction on j at the same time. From (2.38), (2.48), (2.50), we have

$$\begin{cases} J_{12}^{(0)} \equiv -\big((L')^T D_s \overline{A} + (A')^T D_s \overline{L}\big) \mod (\widetilde{T}^0, S^{-1}), \\ J_{21}^{(0)} \equiv -\big((\overline{L}')^T D_s A + (\overline{A}')^T D_s L\big) \mod (\widetilde{T}^0, S^{-1}). \end{cases} \tag{2.55}$$

Since $\overline{\widetilde{N}^+} = \widetilde{N}^-$, $J_{12}^{(0)} \equiv -\bar{J}_{21}^{(0)} \bmod (\widetilde{T}^0, S^{-1})$, (2.32) and (2.55) imply

$$\begin{cases} \Phi_{12}^{(-1)} \equiv -W\big((L')^T D_s \overline{A} + (A')^T D_s \overline{L}\big) \mod (\widetilde{T}^0, S^{-2}), \\ \overline{\Phi_{21}^{(-1)}} \equiv -\Phi_{12}^{(-1)} \mod (\widetilde{T}^0, S^{-2}). \end{cases} \tag{2.56}$$

Now assume that (2.53) and (2.54) are valid for j. From (2.40) and the induction hypothesis, we have

$$J_{12}^{(-j)} \equiv -W^j\big((L')^T D_s^{j+1} \overline{A} + (A')^T D_s^{j+1} \overline{L}\big) \bmod (\widetilde{T}^j, S^{-j-1}). \tag{2.57}$$

Then (2.53) for $j+1$ follows from (2.34) and (2.57).

From (2.34) and $\overline{\widetilde{N}^+} = \widetilde{N}^-$, we deduce

$$\widetilde{N}^+ \overline{\Phi_{21}^{(-j-1)}} - \overline{\Phi_{21}^{(-j-1)}} \widetilde{N}^- - \overline{\sigma(J_{21}^{(-j)})} = 0. \tag{2.58}$$

By using (2.38) and the induction hypothesis, we have

$$\widetilde{N}^+ \overline{\Phi_{21}^{(-j-1)}} - \overline{\Phi_{21}^{(-j-1)}} \widetilde{N}^- - (-1)^{j+1} \sigma(J_{21}^{-j}) \equiv 0 \mod (\widetilde{T}^j, S^{-j-1}). \tag{2.59}$$

Comparing (2.58) with (2.34), we deduce (2.54) for $j+1$.

LEMMA 2.60. *Let*

$$\widetilde{\Phi}_s = \Phi_s - I_{2n}. \tag{2.61}$$

Then we have

(2.62)
$$\begin{cases} \widetilde{\sigma}(R) = -A\left(I_n - A^{-1}[I_n, 0_n]\widetilde{\sigma}(\widetilde{Q}_0\widetilde{\Phi}_0)\begin{bmatrix} I_n \\ 0_n \end{bmatrix}\right) \\ A^{-1}[I_n, 0_n]\widetilde{\sigma}(\widetilde{Q}_0\widetilde{\Phi}_0)\begin{bmatrix} I_n \\ 0_n \end{bmatrix} \simeq \sum_{j=1}^{\infty} A^{-1}[I_n, 0_n]\widetilde{Q}_0\Phi_0^{(-j)}\begin{bmatrix} I_n \\ 0_n \end{bmatrix} \in \sum_{j=0}^{\infty} \tilde{T}^j \cdot S^{-j} \\ \widetilde{\sigma}(-\sqrt{-1}[0_n, I_n]Q_s\widetilde{J}T|_{s=0}) \simeq -\sqrt{-1}L - \sum_{j=1}^{\infty} \sqrt{-1}[0_n, I_n]\widetilde{Q}_0\Phi_0^{(-j)}\begin{bmatrix} I_n \\ 0_n \end{bmatrix} \\ \qquad\qquad \in \sum_{j=0}^{\infty} \tilde{T}^j \cdot S^{-j}. \end{cases}$$

Proof. By the formula for the full symbol of the composition of two pseudodifferential operators, we have

$$\widetilde{\sigma}(\widetilde{Q}_0\widetilde{\Phi}_0) \sim \sum_{k=0}^{\infty}(k!)^{-1}\langle -i\partial_z, \partial_\zeta\rangle^k \left\{\widetilde{Q}_0(y,\zeta)\widetilde{\Phi}_0(z,\eta)\right\}|_{z=y,\ \zeta=\eta}. \tag{2.63}$$

From now on we denote, for convenience, the right-hand side of (2.63) by $\widetilde{Q}_0 \odot \widetilde{\Phi}_0$. By using (2.39) and (2.41) in (2.64), we conclude

$$\begin{aligned} \widetilde{Q}_0 \odot \widetilde{\Phi}_0 &= \sum_{\ell=1}^{\infty} \sum_{j+k=\ell} (k!)^{-1}\langle -i\partial_z, \partial_\zeta\rangle^k \left\{\widetilde{Q}_0(y,\eta)\widetilde{\Phi}_0^{(1j)}(z,\eta)\right\}|_{z=y,\,\zeta=\eta} \\ &\simeq \sum_{\ell=1}^{\infty} \widetilde{Q}_0(y,\eta)\widetilde{\Phi}_0^{(-\ell)}(y,\eta) \in \sum_{\ell=1}^{\infty} \widetilde{T}^\ell \cdot S^{-\ell}. \end{aligned} \tag{2.64}$$

Then (2.62) follows immediately from (2.20) and (2.63).

LEMMA 2.65.

$$\widetilde{\sigma}(R^{-1}) \simeq -A^{-1} - \sum_{j=1}^{\infty} A^{-1}\,[I_n, 0_n]\,\widetilde{Q}_0\Phi_0^{(-j)}\begin{bmatrix} I_n \\ 0_n \end{bmatrix} A^{-1} \in \sum_{j=0}^{\infty} \widetilde{T}^j \cdot S^{-j}. \tag{2.66}$$

Proof. It is easy to prove that $g^{\odot k} \in \sum_{j=0}^{\infty} \widetilde{T}^{j+1} \cdot S^{-j-k}$ for each k $(2 \le k \in \mathbb{N})$ if $g \in \sum_{j=0}^{\infty} \widetilde{T}^{j+1} \cdot S^{-j-1}$. Here $g^{\odot k}$ denotes the kth power of g with respect to the product "$\odot$." Then, by (2.39),

$$\left(A^{-1}[I_n, 0_n]\left(\widetilde{Q}_0 \odot \tilde{\Phi}_0\right)\begin{bmatrix} I_n \\ 0_n \end{bmatrix}\right)^{\odot k} \in \sum_{j=0}^{\infty} \widetilde{T}^{j+1} \cdot S^{-j-k} \tag{2.67}$$

for each $k \in \mathbb{N}$. Hence we get (2.66) from (2.62).

LEMMA 2.68.

$$\widetilde{\sigma}(\Lambda_C) \simeq -Z|\eta| - \sum_{j=1}^{\infty} 2Q^{-1}\bar{A}\Phi_{21}^{(-j)}A^{-1}|\eta|. \tag{2.69}$$

Proof. It is easy to prove that

$$g_1 \odot g_2 \simeq g_1^{(0)} g_2^{(0)} + \sum_{j=1}^{\infty} \left(g_1^{(-j)} g_2^{(0)} + g_1^{(0)} g_2^{(-j)} \right) \in \sum_{j=0}^{\infty} \widetilde{T}^j \cdot S^{-j}$$

for $g_k \simeq \sum_{j=0}^{\infty} g_k^{(-j)} \in \sum_{j=0}^{\infty} \widetilde{T}^j \cdot S^{-j}$ $(k = 1, 2)$. Then from (2.22), (2.62), and (2.66), we conclude

$$\begin{aligned} \widetilde{\sigma}(\Lambda_C) &\simeq \sqrt{-1} L A^{-1} |\eta| \\ &+ \sum_{j=1}^{\infty} \sqrt{-1} \left\{ [0_n, I_n] \widetilde{Q}_0 + L A^{-1} [I_n, 0_n] \widetilde{Q}_0 \right\} \Phi_0^{(-j)} \begin{bmatrix} I_n \\ 0_n \end{bmatrix} A^{-1} |\eta| \\ &= \sqrt{-1} L A^{-1} |\eta| \\ &+ \sum_{j=1}^{\infty} \sqrt{-1} \left[0_n, \bar{L} - L A^{-1} \bar{A} \right] \Phi_0^{(-j)} \begin{bmatrix} I_n \\ 0_n \end{bmatrix} A^{-1} |\eta|. \end{aligned} \tag{2.70}$$

Using the transformation rule for tensors under change of local coordinates, we have from (2.44) and (4.31) in [C-S] (see the Appendix) that

$$Z = -\sqrt{-1}\, L A^{-1} = -Q^{-1} - \sqrt{-1}\, Q^{-1} S. \tag{2.71}$$

Hence, by using (4.32) and (4.33) in [C-S] and the fact that $Z = Z^*$, we conclude

$$\bar{L} - L A^{-1} \bar{A} = 2\sqrt{-1} Q^{-1} \bar{A}. \tag{2.72}$$

By substituting (2.71) into (2.70), we get (2.69).

Now from (2.53), (2.54), and (2.69), we get

$$\widetilde{\sigma}(\Lambda_C) \simeq -Z|\eta| - \sum_{j=1}^{\infty} 2(-1)^{j+1} Q^{-1} \bar{A} \overline{W^j \left((L')^T D_s^j \bar{A} + (A')^T D_s^j \bar{L} \right)} A^{-1} |\eta|. \tag{2.73}$$

By using (2.71), we can replace $(L')^T D_s^j \bar{A} + (A')^T D_s^j \bar{L}$ by $-\sqrt{-1}(A')^T (D_s^j \bar{Z}) \bar{A}$. Thus we obtain

$$\widetilde{\sigma}(\Lambda_C) \simeq -Z|\eta| + \sum_{j=1}^{\infty} 2\sqrt{-1}(-1)^j Q^{-1} \bar{A} \overline{W^j \left((A')^T (D_s^j \bar{Z}) \bar{A} \right)} A^{-1} |\eta|, \tag{2.74}$$

finally proving (2.7).

3. Proofs of the theorems.

(I) *Proof of Theorem* 1.10. Observe that

$$Q = -S_2, \quad S = S_1 \tag{3.1}$$

in terms of the special local coordinates introduced in §2. Then by using (4.52) and (4.53) in the Appendix, we have

$$Z(y_0, \eta^0) = \begin{bmatrix} \frac{2\mu(\lambda+2\mu)}{\lambda+3\mu}, & 0, & \dots & , & 0, & -\sqrt{-1}\frac{2\mu^2}{\lambda+3\mu} \\ & & \mu & & & \\ 0 & & & \ddots & & 0 \\ & & & & \mu & \\ \sqrt{-1}\frac{2\mu^2}{\lambda+3\mu}, & 0, & \dots & , & 0, & \frac{2\mu(\lambda+2\mu)}{\lambda+3\mu} \end{bmatrix} \tag{3.2}$$

in the special local coordinates.

Hence

(3.3)
$$\mu = 2^{-1}\left(Z_{11} + \sqrt{-1}Z_{1n}\right), \quad \lambda = -2^{-1}\sqrt{-1}Z_{1n}^{-1}\left(Z_{11} - 2\sqrt{-1}Z_{1n}\right)\left(Z_{11} + \sqrt{-1}Z_{1n}\right)$$

in these special local coordinates.

Since the choice of these coordinates is stable near y_0 for the fixed η^0, Theorem 1.10 follows directly from Theorem 2.6 and (3.2).

(II) *Proof of Theorem* 1.13.

Let

$$\left(C^{\alpha,\beta}\right) = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

in terms of the local coordinates introduced in §2. For simplicity we set

$$a = \lambda_1,\ b = \lambda_2,\ \lambda_3 = 2^{-1}d \tag{3.4}$$

and assume the following technical condition:

$$a > 0, \quad b > 0, \quad d > 0, \quad ab - d^2 > 0 \quad \text{near } y_0. \tag{3.5}$$

Using (4.32), (4.33) in page 320 of [C-S], we have, after tedious computations, the following formula for

$$Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \tag{3.6}$$

in local coordinates:

$$\begin{cases} Z_{11} &= AB(A+B)\sqrt{A^2+B^2}/\{\sqrt{2}D(A^2+AB+B^2)\} \\ Z_{12} &= -\sqrt{-2}AB\sqrt{A^2+B^2}/\{2(A^2+AB+B^2)\} \\ Z_{21} &= \overline{Z}_{12} \\ Z_{22} &= D(A+B)\sqrt{A^2+B^2}/\{\sqrt{2}(A^2+AB+B^2)\} \end{cases} \tag{3.7}$$

where

$$A = \sqrt{ab - \sqrt{a^2b^2 - abd}},\ B = \sqrt{ab + \sqrt{a^2b^2 - abd^2}}, \quad D = \sqrt{bd}. \tag{3.8}$$

We note that a, b, d can be expressed in terms of A, B, D. By simple algebraic manipulations, (3.7) implies the following formula for A, B, D:

$$\begin{cases} A &= \sqrt{-2}\left(4Z_{11}Z_{22} + Z_{12}^2\right)\left(Z_{11}\sqrt{Z_{22}} \pm \sqrt{Z_{11}^2Z_{22} + Z_{11}Z_{12}^2}\right) \\ &\quad / \left\{Z_{12}\sqrt{4Z_{11}^2Z_{12} + 2Z_{11}Z_{12}^2}\right\}, \\ B &= -\sqrt{-2}\left(4Z_{11}Z_{22} + Z_{12}^2\right)\left(Z_{11}\sqrt{Z_{22}}\right) \pm \sqrt{Z_{11}^2Z_{22} + Z_{11}Z_{12}^2}\Big) \\ &\quad / \left\{Z_{12}\sqrt{4Z_{11}^2Z_{12} + 2Z_{11}Z_{12}^2}\right\}, \\ D &= -\sqrt{-2}\left(4Z_{11}Z_{22} + Z_{12}^2\right)\ \sqrt{-Z_{12}^2Z_{22}} / \left\{Z_{12}\sqrt{4Z_{11}^2Z_{12} + 2Z_{11}Z_{12}^2}\right\}. \end{cases} \tag{3.9}$$

Here the double signs $\pm$ and $\mp$ should read according to the convention that the upper signs and lower signs are grouped together.

Next we show how to recover the derivatives of C at the boundary from the derivatives of Z. By recalling the transformation rule of tensors under change of local coordinates and the stability of the choice of the special local coordinates, we only have to consider (Z_{jk}) and its derivatives.

From now on we simply denote by ∂ any kind of differentiation. Then we have

$$\begin{cases} D^3\partial\left(\frac{Z_{11}}{Z_{22}}\right) = D^{-2}B\partial A + D^{-2}A\partial B - 2D^{-3}AB\partial D \\ D^2\partial\left(-2\sqrt{-1}\frac{Z_{11}}{Z_{12}}\right) = D^{-1}\partial\Lambda + D^{-1}\partial B - D^{-2}(A+B)\partial D \\ (A^2+B^2)^{\frac{1}{2}}(A^2+AB+B^2)^2\partial(\sqrt{-2}Z_{12}) \\ \quad = B^2(A^3+A^2B+B^3)\partial A + A^2(A^3+AB^2+B^3)\partial B. \end{cases} \tag{3.10}$$

The determinant of the coefficients of $\partial A, \partial B, \partial D$ in the right-hand side of (3.10) is $-ABD^{-4}(A^3-B^3)(A^2+B^2) \neq 0$. Hence we can recover $\partial A, \partial B, \partial D$ from $Z_{jk}, \partial Z_{jk}$ $(1 \leq j,k \leq 2)$.

Since the coefficients for $\partial^m A, \partial^m B, \partial^m D \quad (m \geq 2)$ in $\partial^m(\frac{Z_{11}}{Z_{22}})$, $\partial^m(-2\sqrt{-1}\frac{Z_{11}}{Z_{12}})$, $\partial^m(\sqrt{-2}Z_{12})$ are the same as that of $\partial A, \partial B, \partial C$ in $\partial(\frac{Z_{11}}{Z_{22}})$, $\partial(-2\sqrt{-1}\frac{Z_{11}}{Z_{12}})$, $\partial(\sqrt{-2}Z_{12})$, we can also recover $\partial^m A, \partial^m B, \partial^m D \quad (m \geq 2)$ from $\partial^\ell Z_{jk} \quad (1 \leq j,k \leq 2, \quad 0 \leq \ell \leq m)$.

Therefore we have proved Theorem 2.6.

Appendix. In this appendix we point out the modifications which are necessary to generalize Chadwick and Smith's [C-S] results to n-dimensional generic, anisotropic media and n-dimensional isotropic media. We shall write the formulas that need to be changed in [C-S] with the same numbers.

The most important formulas from [C-S] that we have used in the previous sections are (4.31)–(4.33) for n-dimensional, generic, anisotropic media and (4.31)–(4.33), (4.52), (4.53) for n-dimensional, isotropic media.

For n-dimensional, generic, hyperelastic, anisotropic media everything which leads to (4.31)–(4.33) in [C-S] remains basically the same except for the definitions of $Q(\varphi)$ and $R(\varphi)$.

Let $e_j = {}^t(0,\dots,0,1,0,\dots,0) \in \mathbb{C}^n (1 \leq j \leq n)$ and set $m = (\sin\varphi)e_1 - (\cos\varphi)e_n = {}^t(m_1,\dots,m_n), n = (\cos\varphi)e_1 + (\sin\varphi)e_n = {}^t(n_1,\dots,n_n)$. Then we define $Q(\varphi)$ and $R(\varphi)$ by

$$Q(\varphi) = \left(Q_{ij}(n); i \underset{j\to 1,\dots,n}{\downarrow} 1,\dots,n\right), \qquad R(\varphi) = \left(R_{ij}(m,n); i \underset{j\to 1,\dots,n}{\downarrow} 1,\dots,n\right)$$

where

$$Q_{ij}(n) = \sum_{p,q=1}^{n} C_{piqj}n_pn_q, \qquad R_{ij}(m,n) = \sum_{p,q=1}^{n} c_{piqj}m_pn_q.$$

For n-dimensional, isotropic media, we first discuss (4.52) and (4.53) in [C-S] and their related formulas, and then we will discuss (4.31)–(4.33) in [C-S].

The formulas leading to (4.52) and (4.53) need some nontrivial modifications. We define

$$N(0) = \begin{bmatrix} N_1(0), & N_2(0) \\ N_3(0), & N_1^T(0) \end{bmatrix}, \tag{4.45}$$

where

$$N_1(0) = -\begin{bmatrix} \overset{\longleftarrow\ n\ \longrightarrow}{\begin{matrix} 0 & & 1 \\ 0 & \ddots & 0 \\ \frac{\lambda}{\lambda+2\mu} & & 0 \end{matrix}} \end{bmatrix} \begin{matrix} \uparrow \\ n \\ \downarrow \end{matrix}$$

$$N_2(0) = -\begin{bmatrix} \overset{\longleftarrow\ n\ \longrightarrow}{\begin{matrix} \frac{1}{\mu} & & 0 & \\ & \ddots & & \\ & & \frac{1}{\mu} & \\ 0 & & & \frac{1}{\lambda+2\mu} \end{matrix}} \end{bmatrix} \begin{matrix} \uparrow \\ n \\ \downarrow \end{matrix}$$

$$N_3(0) = -\begin{bmatrix} \overset{\longleftarrow\ n\ \longrightarrow}{\begin{matrix} \frac{-4\mu(\lambda+\mu)}{\lambda+2\mu} & & & 0 \\ & -\mu & & \\ 0 & & \ddots & \\ & & & -\mu \\ & & & 0 \end{matrix}} \end{bmatrix} \begin{matrix} \uparrow \\ n \\ \downarrow \end{matrix}$$

$$N(0)\xi_\alpha = i\xi_\alpha \quad (1 \le \alpha \le n-1), \quad N(0)\,\xi_n = i\xi_n + \xi_{n-1}, \tag{4.47}$$

$$\begin{aligned} \xi_\alpha &= \frac{1}{2}(1+i)\mu^{-1/2}\,(0,\dots,0,\overset{\alpha+1}{\underset{\longleftarrow\ n\ \longrightarrow}{\breve{i}}},0,\dots,0,\ 0,\dots,0,\overset{n+\alpha+1}{\underset{\longleftarrow\ n\ \longrightarrow}{\breve{-\mu}}},0,\dots,0), \\ \xi_{n-1} &= \gamma\,(\underset{\longleftarrow\ n\ \longrightarrow}{1,0,\dots,0},\ i,\ \underset{\longleftarrow\ n\ \longrightarrow}{2i\mu,0,\dots,0},-2\mu), \\ \xi_n &= \gamma\,(\underset{\longleftarrow\ n\ \longrightarrow}{-ik,0,\dots,0},\ -k,\ \underset{\longleftarrow\ n\ \longrightarrow}{\mu,0,\dots,0},-i\mu)\,. \end{aligned} \tag{4.48}$$

$$\begin{aligned} N^T(0)\eta_\alpha &= -i\eta_\alpha \quad (1 \le \alpha \le n-2, \alpha = n), \\ N^T(0)\eta_{n-1} &= -i\eta_{n-1} + \eta_n, \\ \eta_{\alpha+n} &= \overline{\eta}_\alpha \quad (1 \le \alpha \le n), \end{aligned} \tag{4.50}$$

$$\begin{aligned} &\eta_\alpha = K\overline{\xi_\alpha} \quad (1 \le \alpha \le n-2), \eta_{n-1} = K\overline{\xi_n},\ \eta_n = K\overline{\xi}_{n-1}, \\ &\eta_{\alpha+n} = K\overline{\xi}_{\alpha+n} \quad (1 \le \alpha \le n-2), \eta_{2n-1} = K\overline{\xi}_{2n},\ \eta_{2n} = K\overline{\xi}_{2n-1}, \end{aligned} \tag{4.51}$$

$$S_1 = -\frac{1}{2}\{(1-2\nu)/(1-\nu)\}\ (e_1 \otimes e_n - e_n \otimes e_1)\,, \tag{4.52}$$

$$S_2 = \mu^{-1}\left[\left\{\frac{1}{4}(3-4\nu)/(1-\nu)\right\}(e_1 \otimes e_1 + e_n \otimes e_n) + \sum_{j=2}^{n-2} e_j \otimes e_j\right]. \tag{4.53}$$

Now in order to discuss (4.32) and (4.33), we set

$$\xi_\pi = \langle a_\pi, \ell_\pi \rangle, \qquad \eta_\pi = \langle \bar{\ell}_\pi, \bar{a}_\pi \rangle \quad (1 \leq \pi \leq 2n).$$

We note that $\bar{a}_\pi$ and $\bar{\ell}_\pi$ are not always the complex conjugates of a_π and ℓ_π. Then it is not so difficult to prove (4.32) and (4.33).

Formula (4.31) can be proved in the same way as in [C-S] (see especially p. 324), because the length of the Jordan chain of our $N(0)$ is also 2.

REFERENCES

[A-N-S] M. AKAMATSU, G. NAKAMURA, AND S. STEINBERG, *Identification of Lamé coefficients from boundary observations*, Inverse Problems, 7 (1991), pp. 335–354.

[C-S] P. CHADWICK AND G. D. SMITH, *Foundations of the theory of surface waves in anisotropic elastic materials*, in Advances in Applied Mechanics, C. H. Yih, ed., vol. 17, Academic Press, New York, 1977, pp. 303–376.

[K-V] R. KOHN AND M. VOGELIUS, *Determining conductivity by boundary measurements*, Comm. Pure. Appl. Math., 38 (1985), pp. 643–667.

[L-U] J. LEE AND G. UHLMANN, *Determining anisotropic real-analytic conductivities by boundary measurements*, Comm. Pure. Appl. Math., 42 (1989), pp. 1097–1112.

[N] G. NAKAMURA, *Existence and propagation of Rayleigh waves and pulses*, in Modern Theory of Anisotropic Elasticity and Applications, J. J. Wu, T. C. T. Ting, and D. M. Barnett eds., Society for Industrial and Applied Mathematics, Philadelphia, 1991, pp. 215–231.

[N-U] G. NAKAMURA AND G. UHLMANN, *Identification of Lamé parameters by boundary observations*, American Journal of Math., 115 (1993), pp. 1161–1187.

[N-UI] ———, *Global uniqueness for an inverse boundary value problem arising in elasticity*, Invent. Math., to appear.

[S-U] J. SYLVESTER AND G. UHLMANN, *Inverse boundary value problems at the boundary-continuous dependence*, Comm. Pure Appl. Math., 41 (1988), pp. 197–219.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 280–304, March 1995

002

LOCAL INVERTIBILITY OF SOBOLEV FUNCTIONS*

I. FONSECA† AND W. GANGBO†

Abstract. A local inverse function theorem is established for mappings $v \in W^{1,N}(\Omega, \mathbb{R}^N)$, $\Omega \subset \mathbb{R}^N$ open set, such that $\det \nabla v(x) > 0$ almost everywhere in $x \in \Omega$. Regularity of the local inverse v^{-1} is obtained provided that $|\frac{\text{adj}(\nabla v)}{\det \nabla v}|^s \det \nabla v \in L^1(\Omega)$ for some $1 \leq s < +\infty$. The local invertibility property is used to study the weak lower semicontinuity of a functional involving variation of the domain.

Key words. local invertibility, topological degree, weak lower semicontinuity

AMS subject classification. 49

1. Introduction. The aim of this paper is to give a simple proof of local invertibility of continuous functions $v \in W^{1,N}(\Omega, \mathbb{R}^N)$, where $\Omega \subset \mathbb{R}^N$ is an open set and $\det \nabla v(x) > 0$ almost everywhere in $x \in \Omega$ (Theorem 3.1). We show that the local inverse function w is $W^{1,1}$ and under suitable hypotheses we improve regularity of w to $W^{1,s}$ for some $s > 1$. Precisely, it is shown that v is locally invertible almost everywhere in the sense that for almost every $x \in \Omega$, there is an open neighborhood D of x and there is a function $w \in W^{1,1}(v(D), D)$ such that $v(D)$ is an open set,

$$v \circ w(y) = y \text{ a.e. } y \in v(D), \tag{1}$$

$$w \circ v(x) = x \text{ a.e. } x \in D, \tag{2}$$

and

$$\nabla w(y) = (\nabla v)^{-1}(w(y)) \text{ a.e. } y \in v(D), \tag{3}$$

where $(\nabla v)^{-1}(w(y))$ is the inverse matrix of $\nabla v(w(y))$. Moreover, if we assume that $|\frac{\text{adj}(\nabla v)}{\det \nabla v}|^s \det \nabla v \in L^1(\Omega)$ for some $1 \leq s < +\infty$, then, as in [Sv], we prove that $w \in W^{1,s}(v(D), D)$. One can then deduce easily that if $\det \nabla v(x) \geq \gamma > 0$ a.e. $x \in \Omega$, $v \in W^{1,q}(\Omega)^N$, and $q \geq N(N-1)$, then $v : D \to v(D)$ and $w : v(D) \to D$ are homeomorphisms, (1) holds for every $y \in v(D)$, (2) holds for every $x \in D$, $w \in W^{1,N}(v(D), D)$, and v is an open mapping on $\Omega \setminus L$ for a suitable $L \subset \mathbb{R}^N$ which has zero measure (see Corollary 3.3). In particular, we conclude that if $N = 2$, $v \in W^{1,2}(\Omega)^2$, and $\det \nabla v(x) \geq \gamma > 0$ a.e. $x \in \Omega$, then $w \in W^{1,2}(v(D), D)$ and there is a set of measure zero $L \subset \mathbb{R}^N$ such that v is an open mapping on $\Omega \setminus L$ and a weaker version of [IS] is obtained. Recently, we became aware of a result by Heinonen and Koskela [HK], where they show that if a mapping is in $W^{1,q}$ for some $q > N(N-1)$, its jacobian is positive almost everywhere and $N \geq 3$, then the mapping is open and discrete, and so $L = \emptyset$.

Conversely, if $v \in W^{1,q}(\Omega)^N$ for some $q > N$, $\det \nabla v(x) \neq 0$ a.e. $x \in \Omega$ and if for almost every $x_0 \in \Omega$ v is locally almost invertible in a neighborhood of x_0 in the sense of (1)–(3), then there are open sets $\Omega_1, \Omega_2 \subset \mathbb{R}^N$ and a set of measure zero $N \subset \mathbb{R}^N$

* Received by the editors May 5, 1993; accepted for publication (in revised form) October 1, 1993. This work was supported by the Army Research Office and the National Foundation through the Center for Nonlinear Analysis at Carnegie Mellon University. The research of the first author was partially supported by the National Science Foundation grant DMS-9201215.

† Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

such that $\Omega = \Omega_1 \cup \Omega_1 \cup N$, $\det \nabla v(x) > 0$ a.e. $x \in \Omega_1$ and $\det \nabla v(x) < 0$ a.e. $x \in \Omega_2$ (see Corollary 3.2).

Note that a homeomorphism $v \in W^{1,\infty}(\Omega)^N$ does not need to satisfy $\det \nabla v(x) \neq 0$ a.e. $x \in \Omega$. Such an example is provided in [MZ] (see Remark 3.4).

The result in this paper is in the same spirit as the work in [Ba], [CN], [Sv], and [TQ]. As far as we know, the existence and the regularity of the local inverse function w is not an immediate consequence of these earlier results where assumptions are placed either on the trace $v|_{\partial\Omega}$ or on $|v(\Omega)|$. By an elementary lemma (Lemma 3.5) and the invertibility result found in [TQ], one can obtain the existence of the local inverse function w and then deduce its regularity. Due to his relaxed assumption $q > N - 1$ (here we have $q \geq N$), Tang used an elaborated method to obtain the existence of an inverse $w \in W^{1,1}_{\text{loc}}$ under the condition introduced by [CN], $\int_\Omega \det \nabla v(x) dx \leq |v(\Omega)|$.

The proof that we present here concerning the local invertibility of v is independent of the work by [Ba], [CN], [Sv], and [TQ], and the method employed relies on basic properties of the degree theory.

In the sequel of this paper, we fix a bounded, open set $\Omega \subset \mathbb{R}^N$ and consider a function $v \in W^{1,q}(\Omega)^N$. We denote by ∇v the gradient of v, i.e., the $N \times N$ matrix of the partial derivatives of v, and by $\operatorname{adj} \nabla v$ the adjugate[1] of ∇v.

As an application of the local invertibility property, we study the weak lower semicontinuity of functionals E of the form

$$E(u, v) = \int_\Omega W\Big(\nabla u(x)(\nabla v(x))^{-1}\Big) dx,$$

defined on the set

$$B_{p,q} = \{(u, v) \in W^{1,p}(\Omega, \mathbb{R}^N) \times W^{1,q}(\Omega, \mathbb{R}^N) : \det \nabla v(x) = 1 \text{ a.e. } x \in \Omega\}$$

where $1 \leq p < +\infty$, $N \leq q \leq +\infty$, $\frac{1}{p} + \frac{N-1}{q} = \frac{1}{r} \leq 1$. When $N = 3$, $\nabla u(x) \cdot (\nabla v(x))^{-1}$ represents the lattice of a neutral elastoplastic change of state of a perfect cubic crystal, u is the elastic deformation, and v corresponds to the slip or plastic deformation. (For details, see Ericksen [Er], Davini and Parry [DP], Fonseca and Parry [FP], and Dacorogna and Fonseca [DF].) We prove that under some convexity and growth assumptions on the function W, E is weakly lower semicontinuous on $B_{p,q}$. If $r > 1$ and $q \neq +\infty$, we rely on the div-curl lemma (see Tartar [Ta]) to prove that

$$\nabla u_\epsilon \cdot (\nabla v_\epsilon)^{-1} \rightharpoonup \nabla u \cdot (\nabla v)^{-1} \text{weakly in } L^r$$

whenever

$$(u_\epsilon, v_\epsilon) \rightharpoonup (u, v) \text{ weakly in } B_{p,q}.$$

We notice, however, that the growth condition of W and the weak lower semicontinuity of E on $B_{p,q}$ do not always imply the existence of the minimum of E on $B_{p,q}$. Indeed, $\{(\nabla u_\epsilon \cdot (\nabla v_\epsilon)^{-1}\}$ being relatively compact in L^r does not imply that $\{(\nabla u_\epsilon)\}$ or $\{(\nabla v_\epsilon)^{-1}\}$ are relatively compact in, respectively, L^p, $L^{q/(N-1)}$ (see [DF], Proposition 4.1).

[1] We recall that if A is a $N \times N$ matrix and $\det A \neq 0$, then $\operatorname{adj} A$ is the $N \times N$ matrix such that $A \cdot \operatorname{adj} A = I_N \det A$. In general, the entries of $\operatorname{adj} A$ are certain homogeneous polynomials of degree $N - 1$ of the entries of A.

The paper is organized as follows: In the second section we fix notation and recall some definitions and well-known properties related to the Brouwer degree. In the third section we prove the local invertibility property of the mappings $v \in W^{1,q}(\Omega, \mathbb{R}^N), q \geq N$, under the condition $\det \nabla v(x) > 0$ a.e. $x \in \Omega$. In view of our applications, in addition we prove that if $v_\epsilon \rightharpoonup v$ weakly in $W^{1,q}$ $q \geq N$, $\det \nabla v(x) > 0$ a.e. $x \in \Omega$ and $\det \nabla v_\epsilon(x) = 1$ a.e. $x \in \Omega$ then, up to a subsequence, v_ϵ and v are, respectively, locally invertible on open sets $D_\epsilon(x)$ and $D(x)$ for almost every $x \in \Omega$, where $D_\epsilon(x)$ and $D(x)$ are neighborhoods of x such that $v_\epsilon(D_\epsilon(x)) = v(D(x))$ does not depend on ϵ. The last section is devoted to the applications, where we obtain the weak lower semicontinuity for a class of functionals E on $B_{p,q}$.

2. Preliminaries. In the sequel we will use the following notation.
For $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$, $|x|$ stands for $(|x_1|^2 + \cdots + |x_N|^2)^{1/2}$ and $|x|_\infty := \max\{|x_1|, \ldots, |x_N|\}$. If $A \subset \mathbb{R}^N$ $|A|$ denotes the Lebesgue measure of A, A^c denotes its complement, $\operatorname{dist}(x, A)$ is defined by $\inf\{|x - y| : y \in A\}$, and $\rho(x, A)$ is given by $\inf\{|x - y|_\infty : y \in A\}$.

If $\Omega \subset \mathbb{R}^N$ is an open set, $v \in W^{1,1}(\Omega)^N$, then ∇v is the $N \times N$ matrix of the distributional derivatives $\frac{\partial v_i}{\partial x_j}$. If, furthermore, $\nabla v \in L^N$, then $\det \nabla v$ is the determinant of ∇v.

We recall some properties of mappings.

LEMMA 2.1. *Let Ω be a bounded, open set in $\mathbb{R}^N$ and $v \in (W^{1,N}_{\text{loc}}(\Omega))^N$ such that $\det \nabla v(x) > 0$ a.e. $x \in \Omega$. Then v is a continuous mapping on Ω. Futhermore, if K is a compact set and V is an open set such that $K \subset V \subset\subset \Omega$, then there is a constant C_N depending only on N, such that*

$$|v(x) - v(y)| \leq M^{\frac{1}{N}} C_N \theta(|x - y|)$$

for every $x, y \in K$ that verify $|x - y| \leq \delta$, where

$$M = \int_V |\nabla v(x)|^N dx,$$

$$\theta(t) = \left(\frac{2}{\log(\frac{2}{t})}\right)^{\frac{1}{N}},$$

and

$$\delta = \min\left\{2, \frac{1}{2}(\operatorname{dist}(K, \mathbb{R}^N \setminus V))^2\right\}.$$

Proof. This lemma is an immediate consequence of Theorem 3.5, p. 294, and Proposition 3.3, p. 292 in [GR] and Theorem 4.4, p. 339 in [Re] (see also [Man]). It can also be shown that, under the above hypotheses, v is a monotonic mapping (see the definition of monotonic mapping below).

DEFINITION 2.2 ([GR]). *Let Ω be a bounded, connected, open set in $\mathbb{R}^N$ and $v \in W^{1,N}(\Omega)^N$. We say that v* is monotonic at the point $x \in \Omega$ *if there is a number $0 < r(x) \leq d(x, \partial\Omega)$ such that for almost every $r \in (0, r(x))$ the pre-image of the intersection of the set $v(B(x,r))$ with the unbounded connected component of $\mathbb{R}^N \setminus v(\partial B(x,r))$ is of measure 0 in $B(x,r)$. We say that v is a monotonic mapping in Ω if v is monotonic at every point $x \in \Omega$.*

We make some remarks on the Brouwer degree theory. For details we refer the reader to [Ll], [Sc].

Let $\Omega \subset \mathbb{R}^N$ be a bounded, open set and let $v : \bar{\Omega} \to \mathbb{R}^N$ be a continuous function. For every $p \in \mathbb{R}^N \setminus v(\partial\Omega)$ the *Brouwer degree* $d(v, \Omega, p)$ *of* v *with respect to* Ω *at* p is a well-defined integer depending only on the boundary values of v. In particular, if $v \in C^1(\bar{\Omega})^N$ and $p \in \mathbb{R}^N \setminus (v(\partial\Omega) \cup v(Z_v))$, we have

$$d(v, \Omega, p) = \sum_{x \in v^{-1}(p)} \operatorname{sign} \det \nabla v(x),$$

where

$$\operatorname{sign} t = \begin{cases} 1 & \text{if } t > 0, \\ -1 & \text{if } t < 0, \end{cases}$$

and $v(Z_v)$ denotes the image of the set $\{x \in \Omega : \det \nabla v(x) = 0\}$.

We give some additional properties of the degree.

PROPOSITION 2.3 ([GR]). *Let* $\Omega \subset \mathbb{R}^N$ *be an open, bounded set,* $v \in C^0(\bar{\Omega})^N$, *and let* $p \in \mathbb{R}^N \setminus v(\partial\Omega)$. *Let* C_p *be the connected component of* $\mathbb{R}^N \setminus v(\partial\Omega)$ *containing* p. *Then we have the following properties:*

(4) $d(v, \Omega, p) = d(u, \Omega, p)$ *if* $u \in C^0(\bar{\Omega})^N$ *and* $|u - v|_\infty < \operatorname{dist}(p, v(\partial\Omega))$,

(5) $d(v, \Omega, p) \neq 0 \Rightarrow \exists\, x \in \Omega$ *such that* $v(x) = p$,

(6) $d(v, \Omega, p) = d(v, \Omega, q) \ \forall\ q \in C_p$,

(7) $d(v, \Omega, p) = d(\phi, \Omega, p)$ *if* $\phi \in C^0(\bar{\Omega})^N$ *and* $\phi = v$ *on* $\partial\Omega$.

Moreover, the degree is invariant under homotopy, i.e.,

(8) $$d(H(\cdot, t), \Omega, p) = d(H(\cdot, 0), \Omega, p),$$

for every homotopy $H \in C^0(\bar{\Omega} \times [0, 1])^N$ *such that* $p \notin H(\partial\Omega, t)$ *for every* $t \in [0, 1]$. *Finally, if* $K \subset \bar{\Omega}$ *is a compact set and* $p \notin v(K)$ *then* (*excision property*)

(9) $$d(v, \Omega, p) = d(v, \Omega \setminus K, p)$$

and if $\Omega = \cup_{i=1}^{+\infty} \Omega_i$, Ω_i *mutually disjoint open sets then* (*decomposition property*)

(10) $$\sum_i d(v, \Omega_i, p) = d(v, \Omega, p).$$

Proof. We refer the reader to [Ll].

LEMMA 2.4. *Let* $\Omega \subset \mathbb{R}^N$ *be a bounded, connected, open set and* $v \in W^{1,N}(\Omega)^N$ *such that* $\det \nabla v(x) > 0$ a.e. $x \in \Omega$. *Let* $f : \mathbb{R}^N \to \mathbb{R}$ *be a measurable function. Then*

(i) *for every measurable set* $E \subset \Omega$, $x \to f \circ v(x)$, *and* $y \to N(v, E, y)$ *are measurable and*

(11) $$\int_E f \circ v(x) |\det \nabla v(x)| dx = \int_{\mathbb{R}^N} N(v, E, y) f(y) dy,$$

where $N(v, E, y)$ *is the cardinality of the elements of the set* $\{x \in E : v(x) = y\}$.

(ii) *If, in addition,* f *is a continuous, bounded function, then for every open set* $U \subset\subset \Omega$ *such that* $|\partial U| = 0$

(12) $$\int_{\bar{U}} f \circ v(x) \det \nabla v(x) dx = \int_{\mathbb{R}^N \setminus v(\partial U)} d(v, U, y) f(y) dy.$$

(iii) *If $U \subset\subset \Omega$ is an open set such that $|\partial U| = 0$ and $p \in \mathbb{R}^N \setminus v(\partial U)$, then*

$$d(v, U, p) = \int_U f(v(x)) \det \nabla v(x) dx \tag{13}$$

for any f nonnegative, continuous real-valued function that satisfies $\int_{\mathbb{R}^N} f(x)dx = 1$, with compact support in V, where V is the connected component of $\mathbb{R}^N \setminus v(\partial U)$ containing p.

Remark 2.5. A function $v : \Omega \to \mathbb{R}^N$ is said to satisfy the N property (Lusin's property) if

$$|v(E)| = 0$$

whenever $E \subset \Omega$ is a measurable set such that $|E| = 0$, and v is said to satisfy the N^{-1} property if

$$|v^{-1}(A)| = 0$$

whenever $A \subset \mathbb{R}^N$ is a measurable set such that $|A| = 0$.

(a) It is known that if $v \in W^{1,N}(\Omega)^N$, $\det \nabla v(x) > 0$ a.e. $x \in \Omega$, then v satisfies the N and the N^{-1} property. (See [GR], pp. 296–297.)

(b) Also, if $v \in W^{1,q}(\Omega)^N$ with $q > N$, then v satisfies the N-property. (For details we refer the reader to [MM].)

Proof of Lemma 2.4. We refer the reader to [GR], Theorem 1.8, p. 280, Theorem 2.6, p. 288, or also to [Sv] for the proof of (11) and (12) in the case where U is a domain.

First we prove that (12) is still valid even if U is not connected and (13) is a by-product of this fact. To achieve this, we remark that by Vitali's covering theorem there are $\{D_i\}$, a countable family of open balls mutually disjoint, and a set N of measure zero such that $(\cup_i D_i) \cap N = \emptyset$ and

$$(\cup_i D_i) \cup N = U.$$

Setting $B = \cup_i D_i$, we have $\cup_i \partial D_i \subset \partial B$. If $y \in \mathbb{R}^N \setminus (v(\partial B) \cup v(\partial U))$, then by the decomposition formula (10)

$$\sum_i \chi_{v(D_i)} d(v, D_i, y) = \sum_i d(v, D_i, y) = d(v, B, y). \tag{14}$$

Let $K = \bar{U} \setminus B$. As K is a compact set and $K \subset \partial U \cup N$, if $y \notin v(K)$ then, by the excision property of degree (9), we obtain

$$d(v, U, y) = d(v, U \setminus K, y) = d(v, B, y). \tag{15}$$

By using the fact that v has the N property (see Remark 2.5), $D_i \subset\subset \Omega$, $|\partial U| = |N| = |\partial D_i| = 0$, by (12), (14), and (15) we obtain

$$\begin{aligned} \int_U f \circ v(x) \det \nabla v(x) dx &= \int_B f \circ v(x) \det \nabla v(x) dx \\ &= \sum_i \int_{D_i} f \circ v(x) \det \nabla v(x) dx \end{aligned}$$

$$= \sum_i \int_{v(D_i)\setminus v(\partial D_i)} d(v, D_i, y) f(y) dy$$
$$= \int_{v(B)\setminus v(\partial B)} d(v, B, y) f(y) dy$$
$$= \int_{v(U)\setminus v(\partial U)} d(v, U, y) f(y) dy,$$

and so we obtain (12).

If, furthermore, $\int_{\mathbb{R}^N} f(x)dx = 1$ and the compact support of f is included V, using (12) and the fact that $d(v, D, \cdot)$ is a constant on V, we conclude that

$$\int_D f \circ v(x) \det \nabla v(x) dx = d(v, D, p). \qquad \square$$

DEFINITION 2.6. *Let $\Omega \subset \mathbb{R}^N$ be an open set, let $v : \Omega \to \mathbb{R}^N$ be a function and $x_0 \in \Omega$.*

1. *We say that v is* differentiable at x_0 *if there is a number $R_0 > 0$, a function $\epsilon : \mathbb{R} \to \mathbb{R}$, and a $N \times N$ matrix $\nabla v(x_0)$ such that*

$$v(x_0 + h) = v(x_0) + \nabla v(x_0)h + |h|\epsilon(|h|)$$

for every $h \in B(0, R_0)$ and $\lim_{t\to 0} \epsilon(t) = 0$. In this case we call $\det \nabla v(x_0)$ the Jacobian of v at x_0.

2. *We say that v is* approximately differentiable at x_0 *if there is a set $A \subset \mathbb{R}$ and a $N \times N$ matrix $\nabla v(x_0)$ such that $\lim_{r\to 0} \frac{|A\cap[0,r]|}{r} = 1$ and*

$$\liminf_{t\to 0, t\in A} \gamma_{x_0}(t) = 0,$$

where

$$\gamma_{x_0}(t) = \sup\left\{ \left| \frac{v(x_0 + tz) - v(x_0)}{t} - \nabla v(x_0) z \right| \ : \ |z| = 1 \right\}.$$

In this case we call $\det \nabla v(x_0)$ the weak Jacobian of v at x_0.

LEMMA 2.7. *Let Ω be a bounded open set in $\mathbb{R}^N$.*

(i) *If $v \in W^{1,N}(\Omega)^N$ is a monotonic mapping, then v is almost everywhere in Ω differentiable.*

(ii) *If $v \in W^{1,q}(\Omega)^N, q > N$, then v is almost everywhere in Ω differentiable.*

(iii) *If $v \in W^{1,q}(\Omega)^N, q > N-1$, then v is almost everywhere in Ω approximately differentiable.*

Proof. We refer the reader to [GR, Thm. 5.4, p. 175], to [Re], and to [MZ].

3. Local invertibility in $W^{1,q}$. We first state the main result of this section (Theorem 3.1) and some of its corollaries.

THEOREM 3.1. *Let $\Omega \subset \mathbb{R}^N$ be a bounded, open set and let $v \in W^{1,N}(\Omega)^N$ be a function such that $\det \nabla v(x) > 0$ a.e. $x \in \Omega$. Then for almost every $x_0 \in \Omega$, v is locally almost invertible in a neighborhood of x_0, in the sense that there exists $r \equiv r(x_0) > 0$, an open set $D \equiv D(x_0) \subset\subset \Omega$, and a function $w : B(y_0, r) \to D$ with $y_0 = v(x_0)$ such that*

$$w \in W^{1,1}(B(y_0, r))^N,$$
$$w \circ v(x) = x \text{ a.e. } x \in D,$$
$$v \circ w(y) = y \text{ a.e. } y \in B(y_0, r),$$
$$\nabla w(y) = (\nabla v)^{-1}(w(y)) \ \text{ a.e. } \ y \in B(y_0, r).$$

If, in addition, $|\frac{\operatorname{adj}(\nabla v)}{\det \nabla v}|^s \det \nabla v \in L^1(\Omega)$ *for some* $1 \leq s < +\infty$ *then* $w \in W^{1,s}(B(y_0, r), D)$.

Before proving Theorem 3.1, we list some of its consequences.

COROLLARY 3.2. *Let* $\Omega \subset \mathbb{R}^N$ *be a bounded, open set,* $q \geq N$, *and* $v \in W^{1,q}(\Omega)^N$ *be a function such that* $\det \nabla v(x) \neq 0$ a.e. $x \in \Omega$.

(a) *Assume that* $\Omega_1, \Omega_2 \subset \mathbb{R}^N$ *are two open sets and* $N \subset \mathbb{R}^N$ *is a set of measure zero such that* $\Omega = \Omega_1 \cup \Omega_1 \cup N$, $\det \nabla v(x) > 0$ *a.e.* $x \in \Omega_1$, *and* $\det \nabla v(x) < 0$ *a.e.* $x \in \Omega_2$. *Then for almost every* $x_0 \in \Omega$ v *is locally almost invertible in a neighborhood of* x_0 *in the sense above.*

(b) *Conversely, if* $q > N$, $v \in W^{1,q}(\Omega)^N$ *and if for almost every* $x_0 \in \Omega$ v *is locally almost invertible in a neighborhood of* x_0, *then there are open sets* $\Omega_1, \Omega_2 \subset \mathbb{R}^N$ *and a null set* $N \subset \mathbb{R}^N$ *such that* $\Omega = \Omega_1 \cup \Omega_2 \cup N$, $\det \nabla v(x) > 0$ *a.e.* $x \in \Omega_1$, *and* $\det \nabla v(x) < 0$ *a.e.* $x \in \Omega_2$.

COROLLARY 3.3. *Let* $q \geq N$, *let* $\Omega \subset \mathbb{R}^N$ *be a bounded, open set and let* $v \in W^{1,q}(\Omega)^N$ *be a function such that* $\det \nabla v(x) = 1$ *a.e.* $x \in \Omega$. *Then the inverse function* w *of Theorem* 3.1 *is such that*

$$w \in W^{1,\frac{q}{N-1}}(v(D))^N.$$

If, in addition, $q \geq N(N-1)$ *then* $w \circ v(x) = x$ *for every* $x \in D$, $v \circ w(y) = y$ *for every* $y \in B(y_0, r)$, v *is a local homeomorphism and* v *is an open mapping on* $\Omega \setminus L$ *for some set* $L \subset \Omega$ *of zero measure. In particular, if* $N = 2$ *then* $N(N-1) = N = 2$ *and* v *is a local homeomorphism at* x_0.

We make some remarks and state some lemmas needed for the proofs of Corollaries 3.2 and 3.3, which will appear at the end of this section.

Remark 3.4.

1. As mentioned in the introduction, it has been proven recently by Heinonen and Koskela [HK, Cor. 1.10] that if a mapping is in $W^{1,q}$ for some $q > N(N-1)$ and if its jacobian is positive and $N \geq 3$, then the mapping is open and discrete and so $L = \emptyset$.

2. Recall that $v \in W^{1,N}(\Omega)^N$ is said to be a *mapping of bounded distortion* (or usually a *quasi-regular mapping*) if $|\nabla v(x)|^N \leq K(\det \nabla v(x))$ for almost every $x \in \Omega$ and for some constant K. It is well known that every mapping of bounded distortion $v \in W^{1,N}(\Omega)^N$ is locally a homeomorphism at almost every point $x_0 \in \Omega$. (See [Re, Thm. 6.6, p. 187].) Moreover, mappings of bounded distortion are open mappings or constant in Ω. (See [Re, Thm. 6.4, p. 184].)

3. Note that even if $v \in C^1(\bar{\Omega})^N$ is such that $\det \nabla v(x) \geq \gamma > 0 \quad \forall x \in \Omega$, we cannot expect a global invertibility of v without any regularity assumptions on the trace of v (see [Ba]).

4. Under the assumptions of Theorem 3.1, we cannot expect v to be locally invertible everywhere (see [Ba]).

5. An example of a mapping $v \in W^{1,\infty}(\Omega)^2$, $(\Omega \subset \mathbb{R}^2)$, is exhibited in [Ba], with $\det \nabla v(x) = 1$ a.e. $x \in \Omega$, for which there is no sequence $v_r \in C^1(\bar{\Omega})^2$ such that $v_r \to v$ uniformly and $J_{v_r}(x) > 0$ a.e. $x \in \Omega$. Therefore, to prove Theorem 3.1, one cannot approximate the function v by a sequence of smooth functions v_r, expecting the functions v_r to be locally invertible.

6. Note that for every bounded, open set $\Omega \subset \mathbb{R}^N$, there exists a measurable set $E \subset \Omega$ of nonzero measure and a homeomorphism $v \in W^{1,\infty}(\Omega)^N$ such that $\det \nabla v(x) = 0$ for every $x \in E$. (See [MZ, Remarks 3.7].)

7. Due to the previous remark, the assumption $\det \nabla v(x) \neq 0$ a.e. $x \in \Omega$ in Corollary 3.2 is essential.

LEMMA 3.5. *Let $\Omega \subset \mathbb{R}^N$ be an open set and let $v \in C^0(\Omega)^N$ and $x_0 \in \Omega$ be such that v is differentiable at x_0. Assume that $\det(\nabla v(x_0)) \neq 0$. Then there is $r_0 > 0$ such that for every $0 < r \leq r_0$ the following assertions hold:*

$$v(x_0 + h) \neq v(x_0) \text{ for every } h \in \bar{B}(0,r) \setminus \{0\}, \tag{16}$$

$$d(v, B(x_0, r), v(x_0)) = \operatorname{sign}(\det \nabla v(x_0)). \tag{17}$$

Proof. We refer the reader to [Re].

Remark 3.6. The relation between differentiability and topological degree was first observed by Reshetnyak [Re].

LEMMA 3.7. *Let $\Omega \subset \mathbb{R}^N$ be an open set and let $v \in W^{1,N}(\Omega)^N$ be such that $\det \nabla v(x) > 0$ a.e. $x \in \Omega$. Then for every $x_0 \in \Omega$ such that v is differentiable at x_0 and $\det \nabla v(x_0) > 0$, there is $R_0 \equiv R_0(x_0)$ such that for every $0 < R < R_0$ the following hold:*

$$N(v, B(x_0, R), y) = 1 \text{ for almost every } y \in C_R, \tag{18}$$

$$d(v, B(x_0, R), y) = 1 \text{ for every } y \in C_R, \tag{19}$$

$$d(v, B, y) = 1 \text{ for every } y \in v(B) \setminus v(\partial B), \tag{20}$$

for every nonempty, open set $B \subset v^{-1}(C_R) \cap B(x_0, R)$ such that $|\partial B| = 0$,

where C_R is the connected component of $\mathbb{R}^N \setminus v(\partial B(x_0, R))$ that contains $y_0 := v(x_0)$ and $N(v, E, y)$ is the cardinality of the set $\{x \in E : \ v(x) = y\}$.

Proof. By Lemmas 2.1 and 2.7, v is continuous and monotonic on Ω and is differentiable at almost every point $x \in \Omega$. Fix $x_0 \in \Omega$ such that v is differentiable at x_0 and $\det \nabla v(x_0) > 0$.

Proof of (19). By Lemma 3.5 there is $R_0 > 0$ such that $B(x_0, R_0) \subset\subset \Omega$ and $d(v, B(x_0, R), y_0) = 1$ for every $0 < R < R_0$ and (19) follows from (6).

Proof of (18). By using the fact that $\det \nabla v(x) > 0$ a.e. $x \in \Omega$, we have that (11), (12), and (19) yield (18).

Proof of (20). Since v satisfies the N property (see Remark 2.5), $|v(\partial B(x_0, R_0))| = 0$ and $|v(\partial B)| = 0$. Since B is a nonempty open set, by (11) we have that $|v(B)| \neq 0$, so $|v(B) \setminus v(\partial B)| \neq 0$. Let $y \in v(B) \setminus v(\partial B)$ and C be the connected component of $\mathbb{R}^N \setminus v(\partial B)$ containing y. As $|v(\partial B(x_0, R_0))| = 0$ and since $d(v, B, \cdot)$ is a constant on C, we may assume without loss of generality that $y \notin v(\partial B(x_0, R_0))$. Let $\rho_\epsilon \in C^\infty(\mathbb{R}^N)$ be such that

$$0 \leq \rho_\epsilon(y), \ \forall y \in \mathbb{R}^N \ \ \forall \epsilon > 0,$$

$$\frac{1}{2} \leq \rho_\epsilon(y) \ \forall y \in B\left(0, \frac{\epsilon}{2}\right), \tag{21}$$

$$\operatorname{supp} \rho_\epsilon \subset B(0, \epsilon) \ \ \forall \epsilon > 0,$$

$$\int_{\mathbb{R}^N} \rho_\epsilon(y) dy = 1 \ \ \forall \epsilon > 0.$$

Since $y \in v(B)$, there is $x \in B$ such that $y = v(x)$. By (6) we have

$$\lim_{\epsilon \to 0} \int_B \rho_\epsilon(v(z) - y) \det \nabla v(z) dz = d(v, B, y) \tag{22}$$

and by using the continuity of v at x, we deduce that for every $\epsilon > 0$ there is $\delta > 0$ such that $|v(z) - y| \leq \frac{\epsilon}{2}$ for every $z \in B(x, \delta)$. By recalling that $\det \nabla v(z) > 0$ a.e. $z \in B(x, \delta)$, by (21) and (22) we obtain

$$d(v, B, y) > 0. \tag{23}$$

Finally, since the degree $d(v, \cdot, y)$ is a nondecreasing function of the set, by using (19) and the fact that $B \subset v^{-1}(C_R) \cap B(x_0, R)$, we obtain

$$d(v, B, y) \leq d(v, B(x_0, R), y) = 1, \tag{24}$$

which, together with (23) and the fact that the degree is an integer number, yields (20). □

LEMMA 3.8. *Let Ω, v, R_0 and x_0 be as in Lemma* 3.7, (18), *and* (19). *Let C_{R_0} be the connected component of $\mathbb{R}^N \setminus v(\partial B(x_0, R_0))$ containing $y_0 := v(x_0)$. Then for every $r > 0$ such that $B(y_0, r) \subset\subset C_{R_0}$, if $O := v^{-1}(B(y_0, r)) \cap B(x_0, R_0) \subset\subset B(x_0, R_0)$ then*

$$v(O) = B(y_0, r), \quad v(\partial O) \subset \partial v(O) = \partial B(y_0, r). \tag{25}$$

Proof. It is clear that $v(O) \subset B(y_0, r)$. Conversely, if $y \in B(y_0, r)$, by (19) $d(v, B(x_0, R_0), y) = 1$ and so by (5) there exists $x \in B(x_0, R_0)$ such that $y = v(x)$, which implies $y \in v(O)$. Let $x \in \partial O$ and let $\{a_n\} \subset O$, $\{b_n\} \subset B(x_0, R_0) \setminus O$ be such that

$$\lim_{n \to +\infty} a_n = \lim_{n \to +\infty} b_n = x.$$

We have $v(a_n) \in v(O) = v(v^{-1}(B(y_0, r))) = B(y_0, r)$ and $v(b_n) \notin v(O) = B(y_0, r)$. By using the continuity of v at x, we have

$$v(x) = \lim_{n \to +\infty} v(a_n) = \lim_{n \to +\infty} v(b_n),$$

which gives $x \in \partial v(O)$. □

LEMMA 3.9. *Let $v \in W^{1,N}(\Omega)^N$, $\det \nabla v(x) > 0$ a.e. $x \in \Omega$ and let $x_0 \in D$ be such that $v(x) \neq v(x_0)$ for every $x \in \bar{B}(x_0, R_0) \setminus \{x_0\}$. Let $0 < R < R_0$ and let C be an open set containing $y_0 = v(x_0)$. Then there is $r > 0$ such that $v^{-1}(B(y_0, r)) \cap B(x_0, R) \subset\subset B(x_0, R)$.*

Proof. Define

$$d(\delta) = \sup\{|x - x_0| : x \in \bar{B}(x_0, R), \ |v(x) - v(x_0)| \leq \delta\}.$$

Since $v(x) \neq v(x_0)$ for every $x \in \bar{B}(x_0, R) \setminus \{x_0\}$ and v is uniformly continuous on $\bar{B}(x_0, R)$, we have

$$\lim_{\delta \to 0} d(\delta) = 0.$$

Take now $r > 0$ such that $d(r) < \frac{R}{2}$. We have

$$v^{-1}(B(y_0, r)) \cap B(x_0, R) \subset B\left(x_0, \frac{R}{2}\right) \subset\subset B(x_0, R). \qquad \square$$

Proof of Theorem 3.1. Let Ω' be the set of points $x_0 \in \Omega$ such that v is differentiable at x_0 and $\det \nabla v(x_0) > 0$. By Lemmas 2.1 and 2.7 we obtain $|\Omega \setminus \Omega'| = 0$. In the sequel, we fix $x_0 \in \Omega'$, set $y_0 = v(x_0)$, and show that v is locally invertible at x_0. By Lemmas 3.5 and 3.7 there is $R_0 > 0$ such that $B(x_0, R_0) \subset\subset \Omega$,

$$N(v, B(x_0, R_0), y) = 1 \text{ a.e. } y \in C_{R_0}, \tag{26}$$

where C_{R_0} is the connected component of $\mathbb{R}^N \setminus v(\partial B(x_0, R_0))$ containing y_0, with $N(v, B(x_0, R_0), y_0) = 1$. By Lemma 3.9 we deduce that there is $r > 0$ such that

$$v^{-1}(B(y_0, r)) \cap B(x_0, R_0) \subset\subset B(x_0, R_0) \tag{27}$$

and

$$B(y_0, r) \subset\subset C_{R_0}. \tag{28}$$

Setting $D = v^{-1}(B(y_0, r)) \cap B(x_0, R_0)$, by (27) and (28) we have $D \subset v^{-1}(C_{R_0}) \cap B(x_0, R_0)$ and by Lemma 3.8

$$v(D) = B(y_0, r), \quad v(\partial D) \subset \partial v(D) = \partial B(y_0, r). \tag{29}$$

By the N^{-1} property of v (see Remark 2.5 and (29)), we have $|\partial D| = 0$, which together with (20) yields

$$d(v, D, y) = 1 \ \ \forall\, y \in v(D) \setminus v(\partial D). \tag{30}$$

By using the definition of D, the fact that $D \subset B(x_0, R_0)$, (26), and (28), we obtain

$$N(v, D, y) = 1 \text{ a.e. } y \in v(D). \tag{31}$$

Let $N := \{y \in v(D) \equiv B(y_0, r) : \ d(v, D, y) \neq 1\}$ and define the candidate for local inverse function, w, by

$$w(y) = x \ \text{ if } \ y \in v(D) \setminus N \ \text{ and } \ v(x) = y, \quad x \in D, \tag{32}$$

$$w(y) = x, \ \text{ if } \ y \in N, \ \ v(x) = y, \tag{33}$$

$x \in D$ being chosen by the axiom of choice.

Claim 1. $w \in L^\infty(B(y_0, r))^N$. We have $w(y) \in D \subset \Omega$ for every $y \in v(D)$ and so w is uniformly bounded in $v(D)$. To prove that w is Borel measurable, fix $\alpha \in \mathbb{R}$ and show that the set

$$A := \{y \in v(D) : \ w_i(y) \geq \alpha\}$$

is measurable. We obtain $A = A_1 \cup A_2$ where

$$A_1 := \{y \in v(D) \setminus N : \ w_i(y) \geq \alpha\},$$

$$A_2 := \{y \in N : \ w_i(y) \geq \alpha\}.$$

Since $|A_2| = 0$, we deduce that A_2 is measurable. By using the fact that the restriction of v to $v^{-1}(v(D) \setminus N)$ is one-to-one, one can see that

$$\begin{aligned} A_1 &= \{v(x) : \ x \in v^{-1}(v(D) \setminus N), \ x_i \geq \alpha\} \\ &= (v(D) \setminus N) \cap \left(\cup_{n=0}^{+\infty} v\{x \in \bar{B}(x_0, R_0) : \alpha + n \leq x_i \leq \alpha + n + 1\} \right). \end{aligned}$$

By using the fact that for every $n \in \mathbb{N}$, $\{x \in \bar{B}(x_0, R_0) : \alpha + n \leq x_i \leq \alpha + n + 1\}$ is a compact set, v is a continuous function, and $v(D) \setminus N$ is measurable, we obtain that A_1 is measurable and we conclude that $w \in L^\infty(B(y_0, r))^N$.

Claim 2.

$$v \circ w(y) = y \text{ for every } y \in v(D) \equiv B(y_0, r), \tag{34}$$

$$w \circ v(x) = x \text{ for every } x \in D \setminus v^{-1}(N). \tag{35}$$

This follows immediately from (32) and (33). One notices that, due to (30) and Remark 2.5, $|v^{-1}(N)| = 0$.

Claim 3. $f \circ w$ is measurable for every $f : D \to \mathbb{R}$ measurable.

We know that every Lebesgue measurable set is a union of a Borel measurable set and a set of measure zero. To show that $f \circ w$ is measurable, by Claim 1 it suffices to show that $w^{-1}(R)$ is measurable for every $R \subset D$ such that $|R| = 0$. Let R be a subset of D such that $|R| = 0$. We have by (34) that

$$w^{-1}(R) \subset v(R),$$

and since $|R| = 0$, by the N property of v, we obtain that $|w^{-1}(R)| = 0$. Thus $w^{-1}(R)$ is measurable.

Let $g : v(D) = B(y_0, r) \to \mathbb{R}$ be defined by

$$g(y) = \frac{|\mathrm{adj} \nabla v(w(y))|}{\det \nabla v(w(y))}.$$

Claim 4. $g \in L^1(v(D))$.

By Claim 3, g is measurable. By Lemma 2.4 and (11), where we set $f = \chi_{v(D)}$ the indicator of the set $v(D)$, and by Claim 2 and (31) we obtain

$$\int_{v(D)} |g(y)| dy = \int_D |g \circ v(x)| \det \nabla v(x) dx = \int_D |\mathrm{adj} \nabla v(x)| dx.$$

Therefore $g \in L^1(v(D))$.

Claim 5. $w \in W^{1,1}(v(D))^N$ and $\nabla w(y) = (\frac{\mathrm{adj} \nabla v(w(y))}{\det \nabla v(w(y))})^T$.

To prove Claim 5, we fix $\phi \in C_0^\infty(v(D))$ and set $K = \mathrm{supp}\phi$. We show that

$$\int_{v(D)} w_\alpha(y) \frac{\partial \phi}{\partial y_j}(y) dy = -\int_{v(D)} \frac{(\mathrm{adj} \nabla v(w(y)))^j_\alpha}{\det \nabla v(w(y))} \phi(y) dy.$$

Set $\delta = \mathrm{dist}(K, \partial v(D)) > 0$. By using the uniform continuity of v on $\bar{D} \subset B(x_0, R_0)$, we choose $\epsilon > 0$ such that

$$|v(x) - v(x')| \leq \frac{\delta}{4} \text{ for every } x, x' \in \bar{D},\ |x - x'| \leq \epsilon. \tag{36}$$

Let $\{v_n\} \subset C^\infty(\bar{D})^N$ be such that

$$v_n \to v \text{ in } C^0(\bar{D})^N \tag{37}$$

and

$$v_n \rightharpoonup v \text{ in } W^{1,N}(D)^N.$$

By (37) we can assume without loss of generality that

$$|v - v_n|_\infty \leq \frac{\delta}{4} \text{ for every } n \in \mathbb{N}. \tag{38}$$

By the fact that $v(\partial D) \subset \partial v(D)$ (see (29)), and by (36) and (38), we have that

$$\operatorname{dist}(x, \partial D) < \epsilon \text{ implies } \phi(v_n(x)) = 0$$

and so

$$\phi \circ v_n \in C_0^\infty(D).$$

In the sequel we denote by A_α^j the component of the j row and the α column of the $N \times N$ matrix A. By (11), (31), (35), and the fact that for every $n \in \mathbb{N}$, $\sum_{\alpha=1}^N \frac{\partial (\operatorname{adj} \nabla v_n)_\alpha^j}{\partial x_\alpha} = 0$ for every $j = 1, \ldots, N$, we have

$$\begin{aligned}
\int_{v(D)} w_\alpha(y) \frac{\partial \phi}{\partial y_j}(y) dy &= \int_D w_\alpha(v(x)) \frac{\partial \phi}{\partial y_j}(v(x)) \det \nabla v(x) dx \\
&= \lim_{n \to +\infty} \int_D x_\alpha \frac{\partial \phi}{\partial y_j}(v_n(x)) \det \nabla v_n(x) dx \\
&= \lim_{n \to +\infty} \int_D x_\alpha \sum_{k=1}^N (\operatorname{adj} \nabla v_n(x))_k^j \frac{\partial}{\partial x_k} \phi(v_n(x)) dx \\
&= - \lim_{n \to +\infty} \int_D (\operatorname{adj} \nabla v_n(x))_\alpha^j \phi(v_n(x)) dx \\
&= - \int_D (\operatorname{adj} \nabla v(x))_\alpha^j \phi(v(x)) dx \\
&= - \int_D \frac{(\operatorname{adj} \nabla v(w \circ v(x)))_\alpha^j}{\det \nabla v(w \circ v(x))} \phi(\circ v(x)) \det \nabla v(x) dx \\
&= - \int_{v(D)} \frac{(\operatorname{adj} \nabla v(w(y)))_\alpha^j}{\det \nabla v(w(y))} \phi(y) dy.
\end{aligned}$$

This equality, together with Claim 4, yields Claim 5.

Claim 6. $\nabla w \in W^{1,s}(v(D))$ if and only if $|g \circ v|^s \det \nabla v \in L^1(v(D))$ for $1 \leq s < +\infty$. Recall that $g(y) = \frac{|\operatorname{adj} \nabla v(w(y))|}{\det \nabla v(w(y))}$ and that $w \in L^\infty(v(D))^N$. Thus $\nabla w \in W^{1,s}(v(D))$ if and only if $\nabla w \in L^s(v(D))$. The result now follows from Claim 5 and (11). □

Remark 3.10. It is possible to show that if $v \in W^{1,q}(\Omega)^N$, $q > N - 1$, $\operatorname{adj} \nabla v \in L^{\frac{N}{N-1}}(\Omega)$, $\det \nabla v(x) > 0$ a.e. in Ω and if v is continuous, then there is local invertibility a.e. in Ω, i.e., for a.e. $x_0 \in \Omega$ there exists $r > 0$ such that $v|_{B(x_0,r)}$ is almost everywhere injective with the inverse $w \in BV_{\text{loc}}(v(B(x_0, r)), \mathbb{R}^N)$ and there exists a set $E \subset v(B(x_0, r))$ such that

$$\begin{aligned}
&E \text{ is an open set of } v(B(x_0, r)), \\
&|v(B(x_0, r) \setminus E| = 0, \\
&w \in W^{1,1}(E, \mathbb{R}^N), \\
&v \circ w(y) = y \text{ a.e. } y \in v(B(x_0, r)), \\
&w \circ v(x) = x \text{ a.e. } x \in B(x_0, r).
\end{aligned}$$

To see this, we recall that by Lemma 2.7 (iii) v is approximatively differentiable a.e. in Ω and by adapting the proof of Lemma 3.5 accordingly, it is possible to show that

$$d(v, B(x_0, r), v(x_0)) = 1$$

for some $r > 0$. Let C_0 be the connected component of $\mathbb{R}^N \setminus v(\partial B(x_0, r))$ which contains $v(x_0)$. Then

$$d(v, B(x_0, r), y) = 1 \tag{39}$$

for every $y \in C_0$, so if we choose $0 < r' < r$ such that

$$B(x_0, r') \subset B(x_0, r) \cap v^{-1}(C_0),$$

then by (39) (and since $\det \nabla v > 0$ a.e.) we have

$$d(v, B(x_0, r'), y) \leq 1$$

for every $y \in \mathbb{R}^N \setminus v(\partial B(x_0, r'))$. It suffices now to use the results in [TQ], (1.3)–(1.5), (2.26), and Theorem 3.7 (i). Note, however, that in [TQ], it is assumed that $\operatorname{adj} \nabla v \in L^r$, $r \geq \frac{q}{q-1}$ and if $N - 1 < q < N$, then $\frac{q}{q-1} > \frac{N}{N-1}$.

As it turns out, [TQ]'s results still hold for $r = \frac{N}{N-1}$ as remarked by [MTY] (see Theorem 5.3 in [MTY]).

Proof of Corollary 3.2.

Proof of (a). We have

$$v \in W^{1,N}(\Omega_1)^N, \quad \det \nabla v(x) > 0 \ \text{ a.e. } x \in \Omega_1$$

and

$$v \in W^{1,N}(\Omega_2)^N, \quad \det \nabla v(x) < 0 \ \text{ a.e. } x \in \Omega_2.$$

It suffices to apply Theorem 3.1 to v and to $R_0 v$ in Ω_2, where R_0 is a constant rotation with $\det R_0 = -1$.

Proof of (b). We now assume that $v \in W^{1,q}(\Omega)^N$, $q > N$, $\det \nabla v(x) \neq 0$ a.e. $x \in \Omega$, and for almost every $x_0 \in \Omega$, v is locally almost injective in a neighborhood of x_0 in the sense that there is an open set $D \equiv D(x_0) \subset\subset \Omega$ and there is a function $w : v(D) \to D$ such that

$$w \circ v(x) = x \ \text{ a.e. } x \in D. \tag{40}$$

By Vitali's covering theorem there is a countable family of nonempty, open, mutually disjoint balls $\{B_i,\ i \in \mathbb{N}\}$ and there is a sequence of functions $w_i : v(\bar{B}_i) \to \Omega$ such that $\bar{B}_i \subset \Omega$ and

$$\begin{aligned} &|\Omega \setminus \cup_{i=1}^{+\infty} B_i| = 0, \\ &w \circ v(x) = x \ \text{ a.e. } x \in B_i. \end{aligned} \tag{41}$$

The task ahead will be to partition B_i into three subsets B_i^1, B_i^2, and N_i such that B_i^1, B_i^2 are two open sets, N_i is a set of measure zero, and

$$\begin{aligned} &\det \nabla v(x) > 0 \ \text{ a.e. } x \in B_i^1, \\ &\det \nabla v(x) < 0 \ \text{ a.e. } x \in B_i^2. \end{aligned}$$

Using the fact that $v \in W^{1,q}(B_i)^N$, $q > N$, by Lemma 2.7 and (41) we deduce that there is a set $A_i \subset \bar{B}_i$ of measure zero such that v is differentiable at every $x \in B_i \setminus A_i$,

$$\begin{aligned} &w \circ v(x) = x \ \text{ for every } \ x \in \bar{B}_i \setminus A_i, \\ &\det \nabla v(x) \neq 0 \ \text{ for every } \ x \in \bar{B}_i \setminus A_i. \end{aligned} \tag{42}$$

Let $\{C^j\}$ be the countable collection of the (open) connected components of $\mathbb{R}^N \setminus v(\partial B_i)$. By Remark 2.5 (a), (b) we have

$$|v^{-1}(v(\partial B_i \cup A_i))| = 0. \tag{43}$$

We have the following claim.

Claim 1. $d(v, B_i, v(x)) = \operatorname{sign} \det \nabla v(x)$ for every $x \in B_i \setminus v^{-1}(v(\partial B_i \cup A_i))$.

Fix $x \in B_i \setminus v^{-1}(v(\partial B_i \cup A_i))$.

Step 1. We prove that $d(v, B(x, r_0), v(x)) = \operatorname{sign} \det \nabla v(x)$ for r_0 small enough. By using the fact that v is differentiable and $\det \nabla v(x) \neq 0$, by Lemma 3.5 we deduce that there is $r_0 > 0$ such that for every $0 < r \leq r_0$ we have

$$d(v, B(x, r), v(x)) = \operatorname{sign} \det \nabla v(x).$$

Step 2. We show that $d(v, B_i, v(x)) = \operatorname{sign} \det \nabla v(x)$. Indeed, by setting $K = \bar{B}_i \setminus B(x, r_0)$, we have that K is a compact set included in $\bar{B}_i$, and by (42) we have $v(x) \notin v(K)$ because $v(x) \notin v(A_i)$. By the excision property of the degree (see Proposition 2.3), we obtain

$$d(v, B_i, v(x)) = d(v, B(x, r_0), v(x)) = \operatorname{sign} \det \nabla v(x).$$

Claim 2. $\operatorname{sign} \det \nabla v(x) = \operatorname{sign} \det \nabla v(x')$ for every $x, x' \in v^{-1}(C^j) \setminus v^{-1}(v(\partial B_i \cup A_i))$. Assume that $x, x' \in v^{-1}(C^j) \setminus v^{-1}(v(\partial B_i \cup A_i))$. Using Claim 1 and the fact that the degree $d(v, B_i, \cdot)$ is constant on each C^j, we obtain that $\operatorname{sign} \det \nabla v(x) = \operatorname{sign} \det \nabla v(x')$.

We now conclude the proof of Corollary 3.2(b). Let $I = \{j \in \mathbb{N} : \det \nabla v(x) > 0 \text{ a.e. } x \in v^{-1}(C^j)\}$ and $J = \{j \in \mathbb{N} : \det \nabla v(x) < 0 \text{ a.e. } x \in v^{-1}(C^j)\}$. Set

$$B_i^1 = \cup_{j \in I} v^{-1}(C^j) \cap B_i,$$

$$B_i^2 = \cup_{j \in J} v^{-1}(C^j) \cap B_i,$$

and

$$N_i = B_i \setminus (B_i^1 \cup B_i^2).$$

Then $B_i = B_i^1 \cup B_i^2 \cup N_i$. Set $\Omega_1 = \cup_i B_i^1$, $\Omega_2 = \cup_i B_i^2$ and $N = \Omega \setminus (\Omega_1 \cup \Omega_2)$, then $|N| = 0$ and Ω_1, Ω_2 have the required properties. □

Proof of Corollary 3.3. To obtain that $w \in W^{1,q/(N-1)}(v(D), D)$ we take $s = \frac{q}{N-1}$ in Theorem 3.1. If $q \geq N(N-1)$, then $w \in W^{1,N}$,

$$\det \nabla w(y) = \frac{1}{\det \nabla v(w(y))} > 0 \ \text{ a.e. } \ y \in v(D)$$

and so, by Lemma 2.1 we deduce that w is continuous. Hence v and w are homeomorphisms and v is an open mapping in Ω' for some $\Omega' \subset \Omega$ open, where $|\Omega \setminus \Omega'| = 0$. □

4. Semicontinuity involving variation of the domain. The variational treatment of crystals with defects leads to the study of functionals of the type

$$E(u,v)=\int_{\Omega} W(\nabla u(x)(\nabla v(x))^{-1})dx,$$

where $\Omega \subset \mathbb{R}^N$ is a reference domain, W is the strain energy density, u is the elastic deformation and v represents the slip (rearrangement) or plastic deformation with $\det(\nabla v(x))=1$ a.e. $x \in \Omega$. The underlying kinematical mode for slightly defective crystals was introduced by Davini [Dav] and later developed by Davini and Parry [DP]. As it turns out, matrices of the form

$$\nabla u(x)(\nabla v(x))^{-1}$$

represent lattice matrices of defect-preserving deformations (neutral deformations) and by taking the viewpoint that equilibria correspond to a variational principle, Fonseca and Parry [FP] studied the structure of some kind of generalized minimizers (the Young measure solutions) for the energy $E(\cdot,\cdot)$. (Related variational problems were also investigated in [DP].)

Using the div-curl lemma, it follows that if $u_n \rightharpoonup u$ in $W^{1,\infty}$ $w*$ and $v_n \rightharpoonup v$ in $W^{1,\infty}$ $w*$, then

$$\nabla u_n(\nabla v_n)^{-1} \rightharpoonup \nabla u(\nabla v)^{-1} \text{ in } L^{\infty} \, w*.$$

Lower semicontinuity and relaxation properties of $E(\cdot,\cdot)$ were addressed only under additional material symmetry assumptions on W. Existence and regularity properties for minimizers of $E(\cdot,\cdot)$ were obtained in [DF]. Following this work, we stress the fact that the direct methods of the calculus of variations fail to apply to this problem, as sequential weak lower semicontinuity of $E(\cdot,\cdot)$ is not sufficient to guarantee the existence of minimizers. Indeed, with $W(F)=|F|^r$, it is shown in [DF] that there are no minimizers in $\{(u,v) \in W^{1,\infty} \times W^{1,\infty} : \ u(x)=x \text{ on } \partial\Omega, \det(\nabla v(x))=1 \text{ a.e.}\}$ if $0<r<N=2$, while for $r>N$ existence is obtained for smooth (u,v) (see Theorem 2.3 in [DF]).

It is clear that if $\{(u_n,v_n)\}$ is a minimizing sequence and if $|\nabla u_n(\nabla v_n)^{-1}|^r$ is bounded in L^1, then

$$\nabla u_n(\nabla v_n)^{-1} \rightharpoonup L \text{ in } L^r, u_n|_{\partial\Omega}=u_0, \det(\nabla v_n)=1 \text{ a.e.}$$

and so if some type of lower semicontinuity prevails, then

$$\int_{\Omega} W(L)dx \leq \liminf \int_{\Omega} W(\nabla u_n(\nabla v_n)^{-1})dx. \tag{44}$$

It would remain to show that L would still have the same structure, precisely

$$L=\nabla u(\nabla v)^{-1},$$

where $u|_{\partial\Omega}=u_0$, $\det(\nabla v)=1$ a.e. Note that (44) is always satisfied if W is a convex function. On the other hand, formally, as $\det(\nabla v)=1$ a.e. and setting $w=u(v^{-1})$, the energy becomes

$$\int_{v(\Omega)} W(\nabla w(y))dy,$$

which is now an energy functional involving variations of the domain. Hence, under this new formulation, quasi convexity seems to be more appropriate than convexity (see [AF], [Ba], [Da], and [Mo]).

Suppose that W is a quasi-convex function, i.e.,

$$W(F) \leq \frac{1}{|Q|} \int_Q W(F + \nabla\phi(x))dx,$$

where $Q = (0,1)^N, \phi \in W_0^{1,\infty}(Q)^N$, and $\nabla u_n(\nabla v_n)^{-1} \rightharpoonup L$ in L^r. Can we say that

$$\int_\Omega W(L) \leq \liminf \int_\Omega W(\nabla u_n(\nabla v_n)^{-1})?$$

As an example, consider

$$W(F) = |F|^2 + |\det(F)|.$$

Although we are unable to answer this question, we prove the following result which is the main theorem of this section.

THEOREM 4.1. *Let $W : M^{N\times N} \to \mathbb{R}$ be a quasi-convex function such that*

$$-C_1(1 + |A|^s) \leq W(A) \leq C_2(1 + |A|^r)$$

for some constants $C_1, C_2 > 0, r > s \geq 1$, $p \geq 1$, $q \geq N$, $\frac{1}{p} + \frac{N-1}{q} = \frac{1}{r}$ ($W \geq 0$ if $r = s = 1$). If $u_n \rightharpoonup u$ in $W^{1,p}(\Omega)^N$, $v_n \rightharpoonup v$ in $W^{1,q}(\Omega)^N$, and $\det(\nabla v_n) = 1$ a.e. in Ω, then

$$\int_\Omega W(\nabla u(\nabla v)^{-1})dx \leq \liminf \int_\Omega W(\nabla u_n(\nabla v_n)^{-1})dx.$$

Before proving Theorem 4.1, we make some remarks.

Remark 4.2.

1. It is clear that if $u \in W^{1,p}$, $v \in W^{1,q}$, and $\det \nabla v = 1$ a.e., then $\nabla u(\nabla v)^{-1} \in L^r$.

2. If $r > 1$, then $s < r$ is a necessary condition as the following counterexample shows. This is an adaptation of an idea of Tartar by Ball and Murat [BM]. Here $r = s = 2 = N, \Omega = (0,1)^2$, $W(F) = \det(F)$, $u_n \rightharpoonup u$ in $H^1(\Omega)$, $v_n(x) = x$ and

$$\int_\Omega \det \nabla u \not\leq \liminf \int_\Omega \det \nabla u_n.$$

3. The growth condition cannot be dropped even if W is polyconvex and non-negative. More precisely, if the relation between p, q, r, and s does not occur, the conclusion of Theorem 4.1 may be false. Indeed, using the example by Malý [Ma] with $q = +\infty, p < N - 1, W(F) = \det F, N = r = s$, we may find $u_n \rightharpoonup u$ in $W^{1,p}$, $u(x) \equiv x$ with $v_n(x) \equiv x$, and

$$\int_\Omega |\det(\nabla u)| > \liminf \int_\Omega |\det(\nabla u_n)|.$$

Moreover, the growth condition prescribed in Theorem 4.1 is the well-known growth condition ensuring weak lower semicontinuity of $E(u, id)$ in $W^{1,p}$ (see [AF] and [Da]).

4. We may ask if these results can be extended to the case $\frac{N^2}{N+1} < q < N$, since, due to Müller's result ([Mu]), if we assume that $\mathrm{Det}\nabla v = 1$ a.e. then $\mathrm{Det}\nabla v = \det \nabla v$ a.e. in Ω.

5. Since lower semicontinuity of the energy is obtained in Theorem 4.1, the question now amounts to showing that one can find a minimizing sequence $\{\nabla u_n(\nabla v_n)^{-1}\}$ where $\{u_n\}$ is bounded in $W^{1,p}$ and $\{v_n\}$ is bounded in $W^{1,q}$. Actually, one only needs to show that there exists a sequence $\{f_n\} \subset W^{1,\infty}(\Omega,\Omega)$ such that $v_n \circ f_n$ is bounded in $W^{1,q}$ and

$$\begin{cases} \det \nabla f_n(x) &= 1 \text{ a.e. } x \in \Omega, \\ f_n(x) &= x \ x \in \partial\Omega. \end{cases}$$

Due to the examples provided in [DF], we know that this may not be possible since the infimum of E may be zero, which may prevent the existence of minimizing sequences bounded in $W^{1,p} \times W^{1,q}$.

As usual in variational problems for which existence of minimizers is not guaranteed (such as variational problems for material that change phase and, here, for slightly defective materials), we focus on the properties of the minimizing sequences rather than study the macroscopic limit of $\nabla u_n(\nabla v_n)^{-1}$.

What follows may help to understand better why boundedness of $\{\nabla u_n(\nabla v_n)^{-1}\}$ may not entail the boundedness of $\{\nabla u_n\}$ and $\{\nabla v_n\}$. Using Theorem 4.1, we show that we may construct a minimizing sequence $\{\nabla u_\epsilon(\nabla v_\epsilon)^{-1}\}$ with $|\nabla u_\epsilon|_p = 0(\frac{1}{\epsilon^\alpha})$, $|\nabla v_\epsilon|_q = 0(\frac{1}{\epsilon^\beta})$, for any $\alpha, \beta > 0$.

Consider the "perturbed" family of variational problems

$$E_\epsilon(u,v) = \int_\Omega W(\nabla u(\nabla v)^{-1})dx + \epsilon^{\alpha p}|\nabla u_\epsilon|_p^p + \epsilon^{\beta q}|\nabla v_\epsilon|_q^q,$$

where $u|_{\partial\Omega} = u_0$, $\det \nabla v = 1$ a.e., $\frac{1}{|\Omega|}\int_\Omega v(x)dx = 0$. Using the direct method of the calculus of variations, Poincaré's inequality, and Theorem 4.1, it follows immediately that there exists $(u_\epsilon, v_\epsilon) \in W^{1,p} \times W^{1,q}$ such that

$$E_\epsilon(u_\epsilon, v_\epsilon) = \inf\{E_\epsilon(u,v) : \ (u,v) \in W^{1,p} \times W^{1,q}, \ \det \nabla v = 1 \text{ a.e.}\}.$$

Then, given an admissible pair (u,v)

$$\begin{aligned} E(u,v) &= \lim_{\epsilon\to 0+} E_\epsilon(u,v) \\ &\geq \limsup_{\epsilon\to 0+} E_\epsilon(u_\epsilon, v_\epsilon) \\ &\geq \limsup_{\epsilon\to 0+} E(u_\epsilon, v_\epsilon), \\ &\geq \inf E. \end{aligned}$$

Doing the same with $\liminf_{\epsilon\to 0+} E(u_\epsilon, v_\epsilon)$ and taking the infimum in (u,v), we conclude that

$$\inf E = \lim_{\epsilon\to 0+} E(u_\epsilon, v_\epsilon)$$

and $|\nabla u_\epsilon|_p = 0(\frac{1}{\epsilon^\alpha})$, $|\nabla v_\epsilon|_q = 0(\frac{1}{\epsilon^\beta})$.

The following two lemmas will be useful to prove Theorem 4.1.

LEMMA 4.3. *Let Ω', Ω be two open sets of $\mathbb{R}^N$ such that $\Omega' \subset\subset \Omega$; let $q \geq N$ and $v, v_n \in W^{1,q}(\Omega)^N$ be such that $\det \nabla v(x) = \det \nabla v_n(x) = 1$ a.e. $x \in \Omega$. Assume*

that $v_n \rightharpoonup v$ *in* $W^{1,q}(\Omega)^N$. *Then there exists a subsequence of* $\{v_n\}$ *(not relabelled) such that for almost every* $x_0 \in \Omega'$, *there exist open sets* $D, D_n \subset \Omega'$ *containing* x_0, *there exist* $n_0 \in \mathbb{N}$, $r_0 \equiv r(x_0) > 0$, $w : B(y_0, r_0) \to D$, $w_n : B(y_0, r_0) \to D_n$ *with* $y_0 = v(x_0)$ *such that for* $n \geq n_0$,

$$
\begin{aligned}
&w_n \circ v_n(x) = x \ \text{ a.e. } x \in \bar{D}_n,\\
&v_n \circ w_n(y) = y \ \text{ for every } y \in \bar{B}(y_0, r_0) \ \text{ and } \ v_n(D_n) = B(y_0, r_0),\\
&w \circ v(x) = x \ \text{ a.e. } x \in \bar{D} \ \text{ and } \ v(x_0) \neq v(x) \ \text{ for } x \in D, x \neq x_0,\\
&v \circ w(y) = y \ \text{ for every } y \in \bar{B}(y_0, r_0) \ \text{ and } \ v(D) = B(y_0, r_0),\\
&w_n, w \in W^{1, \frac{q}{N-1}}.
\end{aligned}
$$

Proof. By using Lemma 2.1 and the Ascoli–Arzela theorem we obtain that, up to a subsequence, v_n converges to v uniformly in $\bar{\Omega}'$. By Lemmas 3.7 and 2.7 for almost every $x_0 \in \Omega'$, there is $R_0 > 0$ such that

$$
\begin{aligned}
&B(x_0, R_0) \subset\subset \Omega',\\
&N(v, B(x_0, R_0), y) = 1 \ \text{ for almost every } \ y \in C_{R_0},\\
&d(v, B(x_0, R_0), y) = 1 \ \text{ for every } \ y \in C_{R_0},\\
&d(v, B, y) = 1 \text{ for every } \ y \in B \setminus v(\partial B),\\
&\text{for every nonempty open set } B \subset v^{-1}(C_{R_0}) \cap B(x_0, R_0) \text{ such that } |v(\partial B)| = 0,
\end{aligned}
$$

where C_{R_0} is the connected component of $\mathbb{R}^N \setminus v(\partial B(x_0, R_0))$ containing $y_0 := v(x_0)$. Since v is differentiable at x_0 and $\det \nabla v(x_0) \neq 0$ we may assume without loss of generality that $N(v, B(x_0, R_0), y_0) = 1$. Fix $0 < \epsilon < d\big(y_0, v(\partial B(x_0, R_0))\big)$ and choose $n_0 \in \mathbb{N}$ such that $|v_n - v|_\infty < \epsilon$. Set

$$
A_\epsilon := \{y \in C_{R_0} : \operatorname{dist}(y, v(\partial B(x_0, R_0))) > \epsilon\}.
$$

It is obvious that A_ϵ is a nonempty open set.

Claim 1. $d(v_n, B(x_0, R_0), y)$ exists and is equal to 1 for every $y \in A_\epsilon$ and every $n \geq n_0$. By Proposition 2.3 (4), together with the fact that $d(v, B(x_0, R_0), y) = 1$ for every $y \in C_{R_0}$, we have

$$
d(v_n, B(x_0, R_0), y) = 1 \tag{45}
$$

for every $y \in A_\epsilon$ and every $n \geq n_0$.

By Lemma 3.9 there is $0 < r_0 < R_0$ such that

$$
B(y_0, r_0) \subset\subset A_\epsilon \ \text{ and } \ v^{-1}(B(y_0, r_0)) \cap B(x_0, R_0) \subset\subset B(x_0, R_0). \tag{46}
$$

Claim 2. We claim that

$$
B(y_0, r_0) \subset\subset C^n_{R_0}, \tag{47}
$$

where $C^n_{R_0}$ is the connected component of $\mathbb{R}^N \setminus v_n(\partial B(x_0, R_0))$ that contains y_0.

We prove first that $A_\epsilon \subset \mathbb{R}^N \setminus v_n(\partial B(x_0, R_0))$. Assume on the contrary that there is $y \in A_\epsilon \cap v_n(\partial B(x_0, R_0))$ and choose $x \in \partial B(x_0, R_0)$ such that $y = v_n(x)$. We would have $|v_n(x) - v(x)| = |y - v(x)| > \epsilon > |v_n - v|_\infty$, which yields a contradiction. Fix $r' > r_0$ such that $\bar{B}(y_0, r') \subset A_\epsilon$. We have that $B(y_0, r')$ is a connected set included in $\mathbb{R}^N \setminus v_n(\partial B(x_0, R_0))$ and containing y_0. We deduce that $B(y_0, r') \subset C^n_{R_0}$ and $B(y_0, r_0) \subset\subset C^n_{R_0}$.

Set $D = v^{-1}(B(y_0, r_0)) \cap B(x_0, R_0) \subset\subset \Omega'$ and $D_n = v_n^{-1}(B(y_0, r_0)) \cap B(x_0, R_0) \subset\subset \Omega'$. By using (45)–(47) and arguments similar to the ones of the proof of Theorem 3.1, together with Corollary 3.3, we deduce that for $n \geq n_0$ there is $w_n : \bar{B}(y_0, r_0) \to \bar{D}_n$, there is $w : \bar{B}(y_0, r_0) \to \bar{D}$ such that

$$
\begin{aligned}
&w_n, w \in W^{1, \frac{q}{N-1}}(B(y_0, r_0))^N,\\
&w_n \circ v_n(x) = x \ \text{ a.e. } x \in \bar{D}_n,\\
&v_n \circ w_n(y) = y \ \text{ a.e. } y \in \bar{B}(y_0, r_0),\\
&w \circ v(x) = x \ \text{ a.e. } x \in \bar{D} \ \text{ and } \ v(x_0) \neq v(x) \ \text{ for } \ x \in \bar{D}, x \neq x_0,\\
&v \circ w(y) = y \ \text{ a.e. } y \in \bar{B}(y_0, r_0).
\end{aligned}
$$

Finally by Lemma 3.8, $v_n(D_n) = v(D) = B(y_0, r_0)$.

Remark 4.4.

1. It follows from the proof above that if the conclusion of Lemma 4.3 holds for $r \equiv r(x_0) > 0$ then it holds also for $0 < r' < r$. Thus, as v is continuous on $\bar{D}$, $v(x) \neq v(x_0)$ for $x \in D$ and $x \neq x_0$, we deduce that

$$\lim_{r \to 0} \max\{|x - x_0| : \ x \in \bar{D}, v(x) \in B(y_0, r_0)\} = 0.$$

2. It is possible to show that $\lim_{n \to +\infty} |D \Delta D_n| = 0$. We divide the proof into two cases.

Claim 1. $\lim_{n \to +\infty} |D \setminus D_n| = 0$.

Let $F_\epsilon = B(y_0, r_0 - \epsilon)$ and $O_\epsilon = v^{-1}(F_\epsilon) \cap D$. We prove first that for each ϵ fixed there exists $n_0 \equiv n_0(\epsilon) \in \mathbb{N}$ such that $n \geq n_0$ implies $O_\epsilon \subset D_n$. Indeed, since $\{v_n\}$ converges to v uniformly, there exists $n_0 \equiv n_0(\epsilon) \in \mathbb{N}$ such that $|v - v_n|_\infty \leq \frac{\epsilon}{2}$ for every $n \geq n_0$. If $x \in O_\epsilon$, we obtain

$$|v_n(x) - y_0| \leq |v(x) - y_0| + |v(x) - v_n(x)| < r_0$$

and so $x \in D_n$. As $\cup_\epsilon O_\epsilon = D$ and the sequence (O_ϵ) is nonincreasing, we have

$$\lim_{\epsilon \to 0} |D \setminus O_\epsilon| = 0$$

which, together with the fact that $|D \setminus D_n| \leq |D \setminus O_\epsilon|$ for $n \geq n_0$, yields Claim 1.

Claim 2. $\lim_{n \to +\infty} |D_n \setminus D| = 0$.

For $\epsilon > 0$, take $n_0 \equiv n_0(\epsilon) \in \mathbb{N}$ such that $|v - v_n|_\infty \leq \frac{\epsilon}{2}$ for every $n \geq n_0$. For $n \geq n_0$, we have

$$\left\{x \in B(x_0, R_0) : r - \frac{\epsilon}{2} \leq |v_n(x) - y_0| < r\right\} \subset \{x \in B(x_0, R_0) : r - \epsilon \leq |v(x) - y_0| < r + \epsilon\}$$

and since v has the N^{-1} property (see Remark 2.5) we obtain

$$|\cap_\epsilon \{x \in B(x_0, R_0) : r - \epsilon \leq |v(x) - y_0| < r + \epsilon\}| = |\{x \in B(x_0, R_0) : |v(x) - y_0| = r\}| = 0.$$

To conclude the proof of Claim 2, it suffices to remark that for $n \geq n_0$ we obtain

$$D_n \setminus D \subset \{x \in B(x_0, R_0) : r - \epsilon \leq |v(x) - y_0| < r + \epsilon\}.$$

LEMMA 4.5. *Let $p \geq 1$, $q \geq N$, $r \geq 1$ be such that $\frac{1}{p} + \frac{N-1}{q} = \frac{1}{r}$. Assume that $\Omega \subset \mathbb{R}^N$ is an open, bounded set, $u_n, u \in W^{1,q}(\Omega)^N$, $u_n \rightharpoonup u$ in $W^{1,p}(\Omega)^N$, $v_n, v \in W^{1,q}\Omega)^N$, $\det \nabla v_n = \det \nabla v = 1$ a.e. in Ω and $v_n \rightharpoonup v$ in $W^{1,q}(\Omega)^N$. Let $x_0 \in \Omega$, and w_n, w be, respectively, the local inverse function of v_n, v, in the open neighborhoods D_n, D of x_0, let $y_0 = v(x_0)$ and $B(y_0, r_0)$ be as in Lemma 4.3 and Remark 4.4. Then the following conditions hold:*

(i) $u_n \circ w_n \in W^{1,r}(B(y_0,r_0))^N$ *and* $\nabla(u_n \circ w_n)(y) = \nabla u_n(w_n(y))(\nabla v_n(w_n(y)))^{-1}$ *a.e.,*

(ii) $u_n \circ w_n \rightharpoonup u \circ w$ *in* $W^{1,r}(B(y_0,r_0))^N$ *if* $r > 1$,

(iii) $u_n \circ w_n \to u \circ w$ *in* $L^1(B(y_0,r_0))^N$ *and* $\{u_n \circ w_n\}$ *is bounded in* $W^{1,1}(B(y_0,r_0))^N$ *if* $r = 1$.

Proof. We recall that by Lemma 4.3 we have

$$w_n, w \in W^{1,\frac{q}{N-1}}(B(y_0,r_0))^N, \quad v(D) = B(y_0,r_0), v_n(D_n) = B(y_0,r_0), \tag{48}$$

$$\nabla w(y) = (\nabla v(w(y)))^{-1}, \nabla w_n(y) = (\nabla v_n(w_n(y)))^{-1} \text{ a.e. } y \in B(y_0,r_0), \tag{49}$$

$$N(v,D,y) = N(v_n,D_n,y) = 1 \text{ a.e. } y \in B(y_0,r_0), \tag{50}$$

$$w \circ v(x) = x \text{ a.e. } x \in D,\ w_n \circ v_n(x) = x \text{ a.e. } x \in D_n. \tag{51}$$

First step. We prove that $u \circ w, u_n \circ w_n \in W^{1,r}(B(y_0,r_0))^N$.

In fact, by the change of variables formula (11), (48)–(51) we have

$$\begin{aligned}\int_{B(y_0,r_0)} |u \circ w(y)|^r dy &= \int_{v(D)} |u \circ w(y)|^r N(v,D,y) dy \\ &= \int_D |u(x)|^r dx < +\infty.\end{aligned}$$

Thus

$$u \circ w,\ u_n \circ w_n \in L^r(B(y_0,r_0))^N.$$

Let $\phi \in C_0^\infty(B(y_0,r_0))$. By (11), (48)–(51), and the fact that each vector row of $\operatorname{adj} \nabla v$ is divergence free, we have

$$\begin{aligned}\int_{B(y_0,r_0)} u_i \circ w(y) \frac{\partial \phi}{\partial y_j} dy &= \int_D u_i(x) \frac{\partial \phi}{\partial y_j} \circ v(x) dx \\ &= -\int_D \sum_{l=1}^N \frac{\partial u_i}{\partial x_l}(x)((\nabla v(x))^{-1})_j^l \phi \circ v(x) dx \\ &= -\int_{B(y_0,r_0)} \sum_{l=1}^N \frac{\partial u_i}{\partial x_l}(w(y))((\nabla v)^{-1} \circ w(y))_j^l \phi(y) dy.\end{aligned}$$

Thus

$$u \circ w \in W^{1,r}(B(y_0,r_0))^N$$

and

$$\nabla u \circ w(y) = \nabla u(w(y))(\nabla v(w(y)))^{-1} \text{ a.e. in } B(y_0,r_0).$$

We have a similar result for $u_n \circ w_n$.

Second step. We conclude that $\{u_n \circ w_n\}$ is bounded in $W^{1,r}(B(y_0,r_0))^N$. Indeed

$$\int_{B(y_0,r_0)} |u_n \circ w_n(y)|^r dy = \int_{D_n} |u_n(x)|^r dx \le \int_\Omega |u_n(x)|^r dx.$$

Since $r \le p$ and $\{u_n\}$ is bounded in $W^{1,p}(\Omega)^N$ we deduce that $\{u_n \circ w_n\}$ is bounded in $L^r(B(y_0,r_0))^N$.

Also

$$\int_{B(y_0,r_0)} |\nabla u_n \circ w_n(y)|^r dy = \int_D |\nabla u_n(x)(\nabla w_n(x))^{-1}|^r dx$$
$$\leq C'\Big[\int_\Omega |\nabla u_n(x)|^p dx\Big]^{\frac{r}{p}}\Big[\int_\Omega |\nabla v_n(x)|^{\frac{q}{N-1}} dx\Big]^{\frac{r(N-1)}{p}} \leq C$$

for some constant C which does not depend on y_0, r, and n. Thus $\{u_n \circ w_n\}$ is bounded in $W^{1,r}(B(y_0,r_0))^N$.

Third step. We prove that, up to a subsequence, $u_n \circ w_n$ converges strongly in $L^1(B(y_0,r_0))$ to $u \circ w$. Let $f \in C(\bar{B}(y_0,r_0))$. By Remark 4.4, $\lim_{n\to+\infty} |D\Delta D_n| = 0$ and so

$$\chi_{D_n}(x) \to \chi_D(x) \text{ a.e. } x \in \Omega.$$

By using the fact that $u_n \rightharpoonup u$ in $W^{1,p}(\Omega)^N$, $v_n \rightharpoonup v$ in $W^{1,q}(\Omega)^N$ and assuming, without loss of generality, that $u_n \to u$ a.e., $v_n \to v$ a.e., we obtain by (11) and the Lebesgue dominated convergence theorem that

$$\lim_{n\to+\infty} \int_{B(y_0,r_0)} u_n \circ w_n(y) f(y) dy = \lim_{n\to+\infty} \int_{D_n} u_n(x) f(v_n(x)) dx$$
$$= \int_D u(x) f(v(x)) dx$$
$$= \int_{B(y_0,r_0)} u \circ w(y) f(y) dy.$$

Therefore $u_n \circ w_n$ converges strongly to $u \circ w$ in measure and by applying the Sobolev imbedding theorem to the bounded sequence $\{u_n \circ w_n\}$ in $W^{1,r}(\Omega)$, we conclude that, up to a subsequence, $u_n \circ w_n$ converges strongly in $L^1(B(y_0,r_0))$ to $u \circ w$.

Fourth step. Using the second and the third step we conclude that $\{\nabla u_n \circ w_n\}$ is bounded in $W^{1,r}(\Omega)^N$,

$$u_n \circ w_n \rightharpoonup u \circ w \text{ in } W^{1,r}(\Omega)^N \text{ if } r > 1,$$

and

$$u_n \circ w_n \to u \circ w \text{ in } L^1(\Omega)^N \text{ if } r = 1. \qquad \square$$

We now give the proof of Theorem 4.1.

Proof of Theorem 4.1. Without loss of generality (and, if necessary, after extracting a subsequence of $\{(u_n, v_n)\}$), we assume that

$$\liminf_{n\to+\infty} \int_\Omega W(\nabla u_n(x)(\nabla v_n(x))^{-1}) dx = \lim_{n\to+\infty} \int_\Omega W(\nabla u_n(x)(\nabla v_n(x))^{-1}) dx < +\infty.$$

Fix $\epsilon > 0$ and let $\Omega_\epsilon \subset\subset \Omega$ be an open set such that $|\Omega \setminus \Omega_\epsilon| < \epsilon$. By Lemma 2.1 and the Ascoli–Arzela theorem, without loss of generality we assume that v_n converges to v uniformly in $\bar{\Omega}_\epsilon$. Set

$$C = \{x \in \Omega_\epsilon : v \text{ is differentiable and almost invertible at } x\},$$

$$A = \{D(x) : x \in C, D(x) \text{ is an open set of } \Omega_\epsilon,\ v(D(x)) \text{ is an open ball}\},$$

and

$$\Omega'_\epsilon = \cup_{D\in A} D.$$

As in the proof of Lemma 3.9, it is easy to see that

$$\inf\{\mathrm{diam} D(x) \ : \ D(x) \in A\} = 0$$

for every $x \in C$. By Lemma 4.3 and Vitali's covering theorem (see [Fe], Theorem 2.8.17, p. 151) there exists $\{x^j \ : \ j \in \mathbb{N}\} \subset \Omega_\epsilon$, $\{D^j \ : \ j \in \mathbb{N}\}$, a family of mutually disjoint, open neighborhoods of, respectively, x_j, and a set of measure zero N such that

$$\Omega_\epsilon = N \cup_{j\in\mathbb{N}} D^j,$$

and $v : D^j \to B(y^j, r^j)$ admits an inverse $w^j \in W^{1,q/(N-1)}(B(y^j, r^j), D^j)$, in the sense of Theorem 3.1, for some $r_j > 0$ and with $y^j = v(x^j)$. Recall that

$$\begin{aligned}
&\nabla(u \circ w^j)(y) = \nabla u(w^j(y))(\nabla v(w^j(y)))^{-1} \ \text{ a.e. } \ y \in B(y^j, r^j),\\
&w^j \circ v(x) = x \ \text{ a.e. } \ x \in D^j,\\
&v \circ w^j(y) = y \ \text{ a.e. } \ y \in B(y^j, r^j),
\end{aligned}$$

and $D^j = v^{-1}(B(y^j, r^j)) \cap B(x^j, R^j)$ for some $R^j > 0$. Fix $k \in \mathbb{N}$. By Lemma 4.5 we obtain, for each $j = 1, \dots, k$ and up to a subsequence, the existence of $w_n^j \in W^{1,q/(N-1)}(B(y^j, r^j))^N$, which is the inverse function of $v_n|_{D_n^j}$ where $D_n^j = v_n^{-1}(B(y^j, r^j)) \cap B(x^j, R^j)$. Recall that $\frac{1}{r} = \frac{1}{p} + \frac{N-1}{q}$ and also

$$\begin{aligned}
&w_n^j \circ v_n(x) = x \ \text{ a.e. } \ x \in D_n^j,\\
&u_n \circ w_n^j \in W^{1,r}(B(y^j, r^j))^N,\\
&\nabla(u_n \circ w_n^j)(y) = \nabla u_n(w_n^j(y))(\nabla v_n(w_n^j(y)))^{-1} \ \text{ a.e.},\\
&u_n \circ w_n^j \rightharpoonup u \circ w^j \ \text{ in } \ W^{1,r}(B(y^j, r^j))^N \ \text{ if } \ r > 1,\\
&u_n \circ w_n^j \to u \circ w^j \ \text{ in } \ L^1(B(y^j, r^j))^N,\\
&\{u_n \circ w_n^j\} \ \text{ is bounded in } \ W^{1,1}(B(y^j, r^j))^N \ \text{ if } \ r = 1,\\
&\lim_{n\to+\infty} |D_n^j \Delta D^j| = 0.
\end{aligned}$$

Fix

$$0 < \eta < \min\{r^j \ : \ j = 1, \dots, k\}.$$

There exists $n(\eta) \in \mathbb{N}$ such that for every $n \geq n(\eta)$ we obtain

$$\max\{|v_n(x) - v(x)| \ : \ x \in \Omega_\epsilon\} < \eta.$$

Since $D^j = v^{-1}(B(y^j, r^j)) \cap B(x^j, R^j)$, we deduce that for every $n \geq n(\eta)$

$$D_n^j(\eta) := D_n^j \cap v_n^{-1}(B(y^j, r^j - \eta)) \subset D^j,$$

and so $D_n^i \cap D_n^j = \emptyset$ if $i \neq j$. Set

$$D^j(\eta) := D^j \cap v^{-1}(B(y^j, r^j - \eta)).$$

We divide the rest of the proof of Theorem 4.1 into two cases.

First case. We assume that $1 = r = \frac{1}{p} + \frac{N-1}{q}$ and that there is a constant C such that $0 \leq W(F) \leq C(1 + |F|)$ for every $F \in M^{N\times N}$. Since $W \geq 0$ and $\{D^j(\eta)\}, \{D_n^j(\eta)\}$ are mutually disjoint for every $n \in \mathbb{N}$, we have by [FM]

$$\begin{aligned}
\int_{\cup_{j=1}^k D^j(\eta)} W(\nabla u(x)(\nabla v)^{-1}(x))dx &= \sum_{j=1}^k \int_{D^j(\eta)} W(\nabla u(x)(\nabla v)^{-1}(x))dx \\
&= \sum_{j=1}^k \int_{B(y^j,r^j-\eta)} W((\nabla u \circ w^j)(y))dy \\
&\leq \sum_{j=1}^k \liminf_{n\to+\infty} \int_{B(y^j,r^j-\eta)} W((\nabla u_n \circ w_n^j)(y))dy \qquad (52) \\
&= \sum_{j=1}^k \liminf_{n\to+\infty} \int_{D_n^j(\eta)} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx \\
&\leq \liminf_{n\to+\infty} \sum_{j=1}^k \int_{D^j} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx \\
&\leq \liminf_{n\to+\infty} \int_\Omega W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx.
\end{aligned}$$

By letting η go to zero, k go to infinity, and ϵ go to zero, we have

$$E(u,v) \leq \liminf_{n\to+\infty} E(u_n, v_n).$$

Second case. We assume that $1 < r = \frac{1}{p} + \frac{N-1}{q}$ and that there are some constants $C_1, C_2 > 0, 1 \leq s \leq r$ such that $-C_1(1 + |F|^s) \leq W(F) \leq C_2(1 + |F|^r)$ for every $F \in M^{N\times N}$. The proof follows as in the first case, where on step (52) we use the lower semicontinuity results of [Da] instead of [FM]. Since $\{\nabla u_n(x)(\nabla v_n)^{-1}(x)\}$ is weakly relatively compact in Ω, we have

$$\begin{aligned}
\int_{\cup_{j=1}^k D^j(\eta)} W(\nabla u(x)(\nabla v)^{-1}(x))dx &= \sum_{j=1}^k \int_{D^j(\eta)} W(\nabla u(x)(\nabla v)^{-1}(x))dx \\
&= \sum_{j=1}^k \int_{B(y^j,r^j-\eta)} W((\nabla u \circ w^j)(y))dy \\
&\leq \sum_{j=1}^k \liminf_{n\to+\infty} \int_{B(y^j,r^j-\eta)} W((\nabla u_n \circ w_n^j)(y))dy \\
&= \sum_{j=1}^k \liminf_{n\to+\infty} \int_{D_n^j(\eta)} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx \\
&= \sum_{j=1}^k \liminf_{n\to+\infty} [\int_{D^j} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx \\
&\quad + \int_{D_n^j(\eta)\setminus D^j} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx
\end{aligned}$$

$$
\begin{aligned}
&- \int_{D^j \setminus D_n^j(\eta)} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx] \\
&\leq \sum_{j=1}^{k} \liminf_{n\to+\infty} \int_{D^j} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx \\
&\quad + C_1 \int_{D^j \Delta D_n^j} (1 + |\nabla u_n(x)(\nabla v_n)^{-1}(x)|^s)dx \\
&\leq \liminf_{n\to+\infty} \sum_{j=1}^{k} \int_{D^j} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx \\
&\leq \liminf_{n\to+\infty} \int_{\Omega} W(\nabla u_n(x)(\nabla v_n)^{-1}(x))dx.
\end{aligned}
$$

By letting η go to zero, k go to infinity, and ϵ go to zero, we conclude that

$$E(u, v) \leq \liminf_{n\to+\infty} E(u_n, v_n). \qquad \square$$

REFERENCES

[AF] E. Acerbi and N. Fusco, *Semicontinuity and relaxation for integrals depending on vector valued functions*, J. Math. Pures Appl., 62 (1983), pp. 371–387.

[Ba] J. M. Ball, *Global invertibility of Sobolev functions and the interpenetration of the matter*, Proc. Roy. Soc. Edinburgh Sect. A, 88 (1988), pp. 315–328.

[BM] J. M. Ball and F. Murat, *$W^{1,p}$ quasiconvexity and variational problems for multiple integrals*, J. Funct. Anal., 58 (1984), pp. 225–253.

[CN] P. G. Ciarlet and J. Nečas, *Injectivity and self contact in non linear elasticity*, Arch. Rat. Mech. Anal., 97 (1987), pp. 171–188.

[Da] B. Dacorogna, *Direct Methods in the Calculus of Variations*, vol. 78, Springer-Verlag, New York, 1987.

[DF] B. Dacorogna and I. Fonseca, *A minimization problem involving variation of the domain*, Comm. Pure Appl. Math., 45 (1992), pp. 871–897.

[Dav] C. Davini, *A proposal for a continuum theory of defective crystals*, Arch. Rational Mech. Anal., 96 (1986), pp. 295–317.

[DP] C. Davini and G. Parry, *On the defect-preserving deformations in crystals*, Internat. J. Plasticity, 5 (1989), p. 295–317.

[Er] J. L. Ericksen, *Twinning of crystals* I, in Metastability and Incompletely Posed Problems, S. Antman et al., eds., IMA Vol Appl. Math 3, Springer-Verlag, New York, 1987, p. 77–96.

[Fe] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[FP] I. Fonseca and G. Parry, *Equilibrium configurations of defective crystals*, Arch. Rational Mech. Anal., 97 (1987), pp. 189–223.

[FM] I. Fonseca and S. Müller, *Quasiconvex integrands and lower semicontinuity in L^1*, SIAM J. Math. Anal, 23 (1992), pp. 1081–1098.

[GR] V. M. Gol'dshtein and Yu. G. Reshetnyak, *Quasiconformal Mappings and Sobolev spaces*, vol. 54, Kluwer Academic Publishers, Dordrecht, Germany, 1990.

[HK] J. Heinonen and P. Koskela, *Sobolev mappings with integrale dilatations*, Arch. Rational Mech. Anal., 125 (1993), pp. 81–97.

[IS] T. Iwaniec and V. Šverák, *On mappings with integrable dilatation*, Proc. Amer. Math. Soc., 118 (1993), pp. 181–188.

[Ll] N. G. Lloyd, *Degree Theory*, Cambridge University Press, Cambridge, UK, 1978.

[Ma] J. Malý, *Weak lower semicontinuity of polyconvex integrals*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 681–691.

[Man] J. Manfredi, *Weakly monotone functions*, J. Geom. Anal., to appear.

[MZ] O. Martio and W. P. Ziemer, *Lusin's condition (N) and mappings with non-negative jacobians*, to appear.

[MM] M. Marcus and V. J. Mizel, *Transformations by functions in Sobolev spaces and lower semicontinuity for parametic variational problems*, Bull. Amer. Math. Soc., 79 (1973), pp. 709–795.

[Mo] C. B. MORREY, *Multiple Integrals in the Calculus of Variations*, 1966, Springer.

[Mu] S. MÜLLER, *A remark on the distributional determinant*, C. R. Acad. Sci. Paris, 311 (1990), pp. 13–17.

[MTY] S. MÜLLER, Q. TANG, AND B. S. YAN, *On a new class of elastic deformations not allowing for cavitation*, Analyse Nonlineaire, 11 (1994), pp. 217–243.

[Re] YU. G. RESHETNYAK, *Space Mappings with Bounded Distortion*, Trans. Math. Monographs, American Mathematical Soceity, Providence, RI, vol. 73, 1989.

[Sc] J. T. SCHWARTZ, *Nonlinear Functional Analysis*, Courant Institute Lecture Notes, New York, 1974.

[Sv] V. ŠVERÁK, *Regularity properties of deformations with finite energy*, Arch. Rational Mech. Anal., 100 (1988), pp. 105–127.

[TQ] Q. TANG, *Almost-everywhere injectivity in nonlinear elasticity*, Proc. Roy. Soc. Edinburgh Sect A., 109 (1988), pp. 79–95.

[Ta] L. TARTAR, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, Heriot-Watt Symposium, vol. 4, R. Knops, ed., Res. Notes in Math. 39, Pitman, San Francisco, CA, 1979, pp. 136–212.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 305–328, March 1995

003

MATHEMATICAL ASPECTS OF THE COMBUSTION OF A SOLID BY A DISTRIBUTED ISOTHERMAL GAS REACTION *

JESUS ILDEFONSO DIAZ† AND IVAR STAKGOLD‡

Abstract. When a diffusing gas reacts isothermally with an immobile solid phase, the resulting equations form a semilinear system consisting of a parabolic partial differential equation for the gas concentration coupled with an ordinary differential equation for the solid concentration. Existence and uniqueness proofs are given which include the important case of nonlipschitzian reaction rates such as those of fractional-power type. Various qualitative features of the solution are studied: approach to the steady state; monotonicity in time; and dependence on initial conditions, on the porosity, and on the geometry.

The relationship between the original problem and the pseudo-steady-state approximation of zero porosity is investigated. When the solid reaction rate is nonlipschitzian, there is a conversion front separating a fully converted region adjacent to the boundary and a partially converted interior core. Estimates are given for the time to full conversion. If the gas reaction rate is nonlipschitzian the gas may not at first fully penetrate the solid. Estimates are given for the time at which full penetration occurs.

Key words. gas-solid reactions, reaction-diffusion, combustion, pseudo-steady state

AMS subject classifications. 35K57, 35R35, 35K50, 35K55

1. Introduction and preliminary results.

Many problems of current interest in chemical engineering and metallurgy involve the interactions of diffusing substances with immobile solid phases (see [1] and [16]).

Here we consider the combustion of a porous solid, known as the *pellet*, as it reacts with a gas diffusing through its pores. The reaction, involving only one species of gas and one of solid, is taken to be simple, irreversible, and isothermal. Structural changes during the reaction are neglected. The state variables are the nondimensional concentrations C of the gas and S of the solid. These concentrations are regarded as continuous functions of time t and of a macroscopic position vector x. Unlike the "shrinking core" model, the reaction is not confined to a thin surface, but is distributed throughout the solid at a rate proportional to the product of a function of C and of a function of S. We assume that the medium can be characterized by effective values of diffusivity and porosity that are independent of position, time, and concentrations. These assumptions can be reconciled with models, such as the Sohn–Szekely model (see [29]), based on a grainlike microstructure for the pellet, with the reaction confined to the surface of the grains.

Mass balances for the solid and gas yield the nondimensional equations

$$S_t = -f(S)g(C) \quad \text{in } (0,\infty)\times\Omega, \tag{1.1}$$

$$\varepsilon C_t - \Delta C = \lambda S_t = -\lambda f(S)g(C) \quad \text{in } (0,\infty)\times\Omega. \tag{1.2}$$

* Received by the editors April 12, 1993; accepted for publication (in revised form) September 21, 1993.

† Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain. This research was partially sponsored by DGICYT (Spain) project PB90/0620.

‡ Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19716. This research was supported by National Science Foundation grant DMS 9113072.

We are therefore dealing with a semilinear system consisting of a parabolic partial differential equation coupled with an ordinary differential equation. Since (1.2) was obtained by dividing the corresponding dimensional equation by the diffusivity, both ε and λ are inversely proportional to the diffusivity. In the problems of interest here, the porosity ε falls in the range (0.01, 0.1), and the Thiele modulus λ in the range (1, 100). The nondimensional reaction rate $f(S)g(C)$ is only defined for $S \geq 0$ and $C \geq 0$ and vanishes when either S or C vanishes. By nondimensionalization we have also made $f(1) = g(1) = 1$. In applications, it is important to consider cases where f and g are only Hölder continuous but not differentiable at $0+$. For instance, successive reactions in which the intermediate steps have special properties may yield an overall reaction rate which is a fractional power of the concentration of one or both of the reactants. Another example is the Sohn–Szekely grain model which, when translated to our variables, leads to $f(S) = S^{2/3}$ and $f(S) = S^{1/2}$ for three- and two-dimensional problems, respectively. With this in mind, we make the following assumptions on f and g:

$$\text{(1.3)} \qquad \begin{cases} f \text{ Hölder continuous on } [0,1], f(0) = 0, f(1) = 1; \\ f(S) \text{ positive, monotone increasing for } S > 0; \\ \text{Same conditions on } g \text{ with } C \text{ replacing } S. \end{cases}$$

The special cases $f(S) = S^m, g(C) = C^p$ play a particularly important role in applications. The nonlipschitz cases $m < 1$ and $p < 1$ yield interesting behavior, such as conversion in finite time, dead cores, and moving fronts [10], [13], [28]. The exponents m and p are known as the *orders* of the solid and gas reactions, respectively.

We take Ω to be a smooth, bounded domain in R^N. As initial and boundary conditions associated with (1.1) and (1.2) we choose

$$\text{(1.4)} \qquad \begin{aligned} &(a)\, S(0,x) = S_0(x), \quad (b)\, C(0,x) = C_0(x), \qquad x \in \Omega, \\ &(c)\, C + \alpha C_\nu = 1, \qquad x \in \partial\Omega, \quad t > 0. \end{aligned}$$

In the rest of the paper we shall assume, at least, that

$$(H_0) \qquad \begin{aligned} &S_0 \in L^\infty(\Omega), \quad C_0 \in H^2(\Omega) \cap L^\infty(\Omega) \quad \text{and} \quad 0 \leq S_0(x) \leq 1, \\ &0 \leq C_0(x) \leq 1 \quad \text{for a.e. } x \in \Omega. \end{aligned}$$

By nondimensionalization we can choose $\|S_0\| = 1, \|C_0\| \leq 1$, where $\| \ \|$ stands for the sup norm. In (1.4), ν is the outward normal derivative and $\alpha \geq 0$ is a constant measuring the boundary resistance to mass transfer from the ambient region where the gas concentration is maintained at a uniform value (which has been taken to be unity through nondimensionalization). The special case $\alpha = 0$ leads to a Dirichlet problem.

The problem (1.1)–(1.4) will be referred to as *problem* (P). We seek a solution $(S(t,x), C(t,x))$ with $S \geq 0$ and $C \geq 0$.

Since ε is small, the term εC_t in (1.2) is often neglected in the chemical engineering literature. This changes (1.2) to an elliptic equation for which no initial condition can be imposed. We are thus facing a singular perturbation whose effect cannot be judged a priori. The new problem is known as the *pseudo-steady-state* (p.s.s.) problem, which we denote by $(\hat{\mathrm{P}})$:

$$\hat{S}_t = -f(\hat{S})g(\hat{C}), \tag{1.5}$$

$$-\Delta\hat{C} = \lambda\hat{S}_t = -\lambda f(\hat{S})g(\hat{C}), \tag{1.6}$$

$$\hat{S}(0,x) = \hat{S}_0(x) \geq 0, \qquad \|\hat{S}_0\| = 1, \tag{1.7}$$

$$\hat{C} + \alpha\hat{C}_\nu = 1, \quad x \in \partial\Omega, \quad t > 0. \tag{1.8}$$

The same nondimensionalization used to obtain (P) gives ($\hat{\text{P}}$). No initial condition is imposed on $\hat{C}$ since $u(x) \doteq \hat{C}(x,0)$ is determined as the unique, necessarily nonnegative, solution of the elliptic problem

$$-\Delta u = -\lambda f(\hat{S}_0)g(u), \quad x \in \Omega; \quad u + \alpha u_\nu = 1, \quad x \in \partial\Omega. \tag{1.9}$$

In previous papers [25], [27], it has been shown that the solution of ($\hat{\text{P}}$) provides a reasonable approximation to the solution of (P) when ε is small. Specifically, the following result was obtained for the case $f(S) = S^m$ and $g(C) = C^p$:

$$\int_0^t (\hat{C} - C)\, d\tau \leq \varepsilon\|w\|, \tag{1.10}$$

where $w(x)$ is the solution of the simple Poisson problem

$$-\Delta w = 1, \quad x \in \Omega; \quad w + \alpha w_\nu = 0, \quad x \in \partial\Omega. \tag{1.11}$$

For $\alpha = 0$, this is the so-called *torsion* problem about which a great deal of information is available (see, for instance, Bandle [2] and McNabb and Keady [20]). The uniformity of (1.10) in time is perhaps unexpected because $C(0,x)$ and $\hat{C}(0,x)$ are not within $0(\varepsilon)$ of each other. In §3 of the present paper we shall extend (1.10) to general f and g obeying (1.3).

It is of particular interest to know if the solid is fully converted in finite time. We observe from (1.1) that $S(t,.)$ is monotonically decreasing; if for some point $x = \xi$, we have $S(T,\xi) = 0$, then $S(t,\xi) = 0$ for $t \geq T$. This property does not hold for C since diffusion from neighboring points may raise the gas concentration. If $\alpha = 1, S$ on the boundary obeys the ordinary differential equation $S_t = -f(S)$, which can be explicitly integrated. If the integral

$$\int_0^1 \frac{ds}{f(s)} \doteq I \tag{1.12}$$

is finite (as is the case when $f(S) = S^m$ with $m < 1$) we find that S vanishes on the boundary for $t \geq I$. It then turns out that $S(t,x)$ is identically zero in Ω for all t sufficiently large. The infimum of such times is the time t_1 to full conversion. We shall estimate that time in terms of the corresponding quantity $\hat{t}_1$ for the pseudo-steady-state problem. Section 4 is devoted to this and related questions.

To prove existence of a solution to (P) and ($\hat{\text{P}}$) we first reformulate the problem in §2 by introducing the new variable $X = 1 - S$, which now places the problem in a quasi-monotone framework. A number of authors have used quasi-monotone methods for reaction-diffusion systems (see, for instance, [18] and [22]). Because of the nonlipschitzian character of our nonlinearity and the absence of diffusion in one of the equations, the existence proofs in the literature are not directly applicable to

our problem. Our proof is based on a constructive *nonlinear* iteration scheme which preserves qualitative properties at each step.

We first prove existence by the method just described. Continuous dependence and uniqueness are then proved by an L^1 technique typical of degenerate quasilinear parabolic problems. Some remarks are made about weakening the regularity assumptions through more abstract approaches. The results in this section were already sketched out in our paper [10].

In §3, we consider the asymptotic behavior of the solution as $t \to \infty$ and the relation between (P) and ($\hat{\text{P}}$) when ε is small. We show that $S, \hat{S}$ tend monotonically to zero as $t \to \infty$ and that $C, \hat{C}$ tend to 1. The approach to 1 is monotonic for $\hat{C}$ and will be monotonic for C if C_0 obeys a certain natural condition. The dependence on ε is discussed in a number of theorems. If $(S_\varepsilon, C_\varepsilon)$ is the solution of (P) for $\varepsilon > 0$ and $(\hat{S}, \hat{C})$ the solution of ($\hat{\text{P}}$), we show that, under expected conditions on C_0 and S_0, we have monotonic convergence of S_ε to $\hat{S}$ and of C_ε to $\hat{C}$ as $\varepsilon \to 0$. Other theorems in this section deal with the behavior as $\varepsilon \to 0$ of $\int_\Omega |S - S_\varepsilon|\, dx$ and $\int_0^t (\hat{C} - C_\varepsilon)\, d\tau$. These theorems provide a strong generalization of (1.10).

In §4, we discuss the conversion of the solid and, to a lesser extent, the penetration of the gas. For many practical purposes the quantity of interest is the fraction of solid converted up to time t,

$$\gamma(t) = 1 - \frac{\int_\Omega S(t,x)\, dx}{\int_\Omega S_0(x)\, dx},$$

with a similar definition for $\hat{\gamma}(t)$. Both of these increase monotonically to 1 as $t \to \infty$. Estimates are given for $\gamma(t)$, particularly in the case of full conversion in finite time (I finite in (1.12)) when the quantities of principal interest are the times t_1 and $\hat{t}_1$ to full conversion. Comparison between different types of reaction is also considered.

For certain $g(C)$, for instance if $g(C) = C^p$ with $p < 1$, the gas may not fully penetrate the solid for small t. It is easy to see that this "dead core" must disappear in finite time so that $C, \hat{C}$ are strictly positive for $t \geq T$. We obtain estimates for this dead core as well as for T.

2. Existence, uniqueness, and continuous dependence. In order to use quasi-monotone methods in their simplest form, we begin by replacing $S(t,x)$ by

$$X(t,x) = 1 - S(t,x). \tag{2.1}$$

Since $S(t,\cdot)$ is monotonically decreasing, $X(t,\cdot)$ is monotonically increasing. Note that if $S_0(x) \equiv 1$, as is often the case in applications, then X is the local *fraction of solid* converted by time t.

Problem (P), considered on a finite interval $(0,T)$ then becomes the problem (P′):

$$X_t = F(X)g(C) \quad \text{in } Q_T, \tag{2.2}$$

$$\varepsilon C_t - \Delta C = -\lambda F(X)g(C) = -\lambda X_t \quad \text{in } Q_T, \tag{2.3}$$

$$X(0,x) = 1 - S_0(x), \quad C(0,x) = C_0(x) \quad \text{on } \Omega, \tag{2.4}$$

$$C + \alpha C_\nu = 1 \quad \text{on } \Sigma_T, \tag{2.5}$$

where $Q_T = (0,T)\times\Omega$, $\Sigma_T = (0,T)\times\partial\Omega$, and $F(X) \equiv f(1-X)$ is monotone decreasing in X with $F(0) = 1, F(1) = 0$.

Similar considerations apply to the p.s.s. problem ($\hat{\text{P}}$); see (1.5)–(1.8). Setting $\hat{X} = 1 - \hat{S}$, we obtain problem ($\hat{\text{P}}'$):

$$\hat{X}_t = F(\hat{X})g(\hat{C}) \quad \text{in } Q_T, \tag{2.2a}$$

$$-\Delta \hat{C} = -\lambda F(\hat{X}) g(\hat{C}) = -\lambda \hat{X}_t \quad \text{in } Q_T, \tag{2.3a}$$

$$\hat{X}(0,x) = 1 - \hat{S}_0(x) \quad \text{on } \Omega, \tag{2.4a}$$

$$\hat{C} + \alpha \hat{C}_\nu = 1 \quad \text{on } \Sigma_T. \tag{2.4b}$$

If (2.2), (2.3) is regarded as a system for the vector (X, C), the forcing term $(F(X)g(C), -\lambda F(X)g(C))$ is nondecreasing in the off-diagonal variables, i.e., *quasi-monotone.* The system is then in a form which makes it relatively simple to use the notions of sub- and supersolutions.

DEFINITION 2.1. *Let*

$$\overline{X} \in W^{1,\infty}(0,T : L^\infty(\Omega)) \quad \textit{and} \quad \bar{C} \in H^1(0,T : L^2(\Omega)) \cap L^2(0,T : H^2(\Omega)) \cap L^\infty((0,T) \times \Omega).$$

The pair $(\bar{X}, \bar{C})$ *is said to be a* supersolution *to* (P′) *if*

$$\begin{aligned}
&\bar{X}_t \geq F(\bar{X}) g(\bar{C}) \quad \textit{in } Q_T,\\
&\varepsilon \bar{C}_t - \Delta \bar{C} \geq -\lambda F(\bar{X}) g(\bar{C}) \quad \textit{in } Q_T,\\
&\bar{C} + \alpha \bar{C}_\nu \geq 1 \quad \textit{on } \Sigma_T,\\
&\bar{X}(0,x) \geq X_0(x) \doteq 1 - S_0(x) \quad \textit{in } \Omega,\\
&\bar{C}(0,x) \geq C_0(x) \quad \textit{in } \Omega,
\end{aligned}$$

at almost every point of the corresponding domain. A subsolution $(\underline{X}, \underline{C})$ *satisfies the same conditions with all five inequalities reversed. If* (X, C) *is both a* supersolution *and a subsolution we say that* (X, C) *is a* solution.

We observe that $(0,0)$ is a subsolution and $(1,1)$ is a supersolution.

Next, we introduce the following iteration scheme: given a pair of smooth functions (X^{k-1}, C^{k-1}), $k \geq 2$, we define the pair (X^k, C^k) as the solution of the *uncoupled nonlinear equations*

$$\begin{aligned}
&X^k_t = F(X^k) g(C^{k-1}) \quad \text{in } Q_T,\\
&X^k(0,x) = X_0(x) \quad \text{in } \Omega,\\
&\varepsilon C^k_t - \Delta C^k = -\lambda F(X^{k-1}) g(C^k) \quad \text{in } Q_T,\\
&C^k + \alpha C^k_\nu = 1 \text{ on } \Sigma_T, \quad C^k(0,x) = C_0(x) \quad \text{in } \Omega.
\end{aligned}$$

The existence and uniqueness of the solutions X^k, C^k of these uncoupled problems satisfying

$$X^k \in W^{1,\infty}(0,T : L^\infty(\Omega)) \quad \text{and} \quad C^k \in H^1(0,T : L^2(\Omega)) \cap L^2(0,T : H^2(\Omega)) \cap L^\infty((0,T) \times \Omega)$$

can be found (for instance) in Vrabie [31] (see Theorem 3.10.1).

We are now in a position to prove existence for the equivalent problems (P′) and (P).

THEOREM 2.1. *Assume* (1.3) *and* (H_0). *Let* $X_0 = 1 - S_0$. *Let* $(\underline{X}, \underline{C})$ *be a subsolution and* $(\bar{X}, \bar{C})$ *a supersolution with* $(\underline{X}, \underline{C}) \leq (\bar{X}, \bar{C})$. *Then there exists a solution* (X, C) *of* (P′) *satisfying* $(\underline{X}, \underline{C}) \leq (X, C) \leq (\bar{X}, \bar{C})$.

Proof. Consider the sequences $(\underline{X}^k, \underline{C}^k)$ and $(\bar{X}^k, \bar{C}^k)$ obtained by applying our iteration scheme to $(\underline{X}^1, \underline{C}^1) = (\underline{X}, \underline{C})$ and $(\bar{X}^1, \bar{C}^1) = (\bar{X}, \bar{C})$, respectively. By

repeated application of the comparison principle for the uncoupled equations, we see that the sequence $(\underline{X}^k, \underline{C}^k)$ is monotonically increasing while $(\bar{X}^k, \bar{C}^k)$ is monotonically decreasing. From the hypothesis $(\underline{X}, \underline{C}) \leq (\bar{X}, \bar{C})$ we can also show that $(\underline{X}^k, \underline{C}^k) \leq (\bar{X}^k, \bar{C}^k)$. The monotonically increasing sequence $(\underline{X}^k, \underline{C}^k)$ is bounded above and must therefore converge as stated to $(\underset{\sim}{X}, \underset{\sim}{C})$. To show that $(\underset{\sim}{X}, \underset{\sim}{C})$ is a solution it is enough to use the following a priori estimates:

$$\|\underline{X}_t^k\|_{L^\infty((0,T)\times\Omega)} \leq 1,$$
$$\|\varepsilon \underline{C}_T^k\|_{L^2(0,T:L^2(\Omega))} \leq \|\lambda F(\underline{X}^{k-1}) g(\underline{C}^k)\|_{L^2(0,T:L^2(\Omega))} + \|C_0\|_{H^1(\Omega)}^{1/2} \leq M(T,\Omega)$$

for some constant $M(T,\Omega) > 0$ independent of k (this follows from a well-known result due to Brezis; see Theorem 1.9.3 in Vrabie [31]); and

$$\|\Delta \underline{C}^k\|_{L^2(0,T:L^2(\Omega))} \leq M(T,\Omega) + |\Omega|^{1/2}.$$

Standard arguments show that $(\underset{\sim}{X}, \underset{\sim}{C})$ is a solution with the regularity mentioned in Definition 2.1. Similarly $(\bar{X}^k, \bar{C}^k)$ converges downwards to a solution $(\tilde{X}, \tilde{C})$; clearly $(\tilde{X}, \tilde{C}) \geq (\underset{\sim}{X}, \underset{\sim}{C})$. □

Remark 2.1. Under the additional assumptions

$$C_0 \in C^{2+\delta}(\Omega) \quad \text{and} \quad S_0 \in C^{\delta}(\Omega),$$

it is possible to show that the functions $(\underline{X}, \underline{C}), (\bar{X}, \bar{C})$ obtained in Theorem 2.1 are, in fact, classical solutions. Indeed, we point out first that the (unique) solutions (X^k, C^k) of the uncoupled problems are classical solutions as follows from an existence result due to Pao [21]. In order to show that the limit $(\underset{\sim}{X}, \underset{\sim}{C})$ is a classical solution we can proceed as follows: since $\varepsilon(\underset{\sim}{C}_t - \Delta \underset{\sim}{C}) \in L^\infty((0,T)\times\Omega)$ we deduce by well-known regularity results (see, e.g., references in [15]) that $\underset{\sim}{C} \in C^\mu([0,T]\times\bar{\Omega})$. Thus $g(\underset{\sim}{C})$ is a Hölder continuous function. Using the explicit formula (4.4) for $\underset{\sim}{X}$ we conclude that $\underset{\sim}{X}$, and hence $F(\underset{\sim}{X})$, is a Hölder continuous function. Finally, from the equation, $\varepsilon(\underset{\sim}{C}_t - \Delta \underset{\sim}{C})$ is a Hölder continuous function, which implies that $\underset{\sim}{C} \in C^{1,2}_{t,x}((0,T)\times\bar{\Omega}) \cap C^0([0,T]\times\bar{\Omega})$ and satisfies the equation in the classical sense.

Remark 2.2. The existence of solutions for the coupled system can also be obtained by means of fixed point arguments using the compactness of some suitable "Green operator." This approach is developed in the article by Diaz and Vrabie [11], where the case of f and g discontinuous at the origin is also considered.

The existence proof of Theorem 2.1 constructs two solutions $(\underset{\sim}{X}, \underset{\sim}{C})$ and $(\tilde{X}, \tilde{C})$. The following theorem, using an L^1 technique typical of some degenerate quasilinear parabolic equations, proves continuous dependence and uniqueness for a general class of solutions (including the ones obtained in Theorem 2.1).

THEOREM 2.2. *Let f and g be continuous nondecreasing functions with $f(0) = g(0) = 0$. Let $(S,C), (S^*,C^*)$ be solutions of* (P) *(in the sense of Definition* 2.1*) corresponding to the initial data $(S_0,C_0), (S_0^*,C_0^*)$. Then for any $t \geq 0$ we have that*

$$\begin{aligned}(2.6)\qquad &\varepsilon \int_\Omega |C(t,x) - C^*(t,x)|\,dx + \lambda \int_\Omega |S(t,x) - S^*(t,x)|\,dx \\ &\quad + \frac{1}{\alpha}\int_0^t \int_{\partial\Omega} |C(\tau,\sigma) - C^*(\tau,\sigma)|\,d\tau\,d\sigma \\ &\leq \varepsilon \int_\Omega |C_0(x) - C_0^*(x)|\,dx + \lambda \int_\Omega |S_0(x) - S_0^*(x)|\,dx.\end{aligned}$$

In particular, a solution (S, C) *(in the sense of Definition* 2.1*) is unique.*

Proof. We start by assuming f and g strictly increasing. We have that

$$(S - S^*)_t + (f(S) - f(S^*))g(C^*) = -f(S)(g(C) - g(C^*)), \tag{2.7}$$

$$\varepsilon(C - C^*)_t - \Delta(C - C^*) + \lambda f(S)(g(C) - g(C^*)) = -\lambda(f(S) - f(S^*))g(C^*). \tag{2.8}$$

Now, multiplying (2.7) by sign $(S - S^*)$ and using

$$\int_0^t \int_\Omega h_t(\tau, x)\text{sign}\,(h(\tau, x))\,dx\,d\tau = \int_\Omega |h(t, x)|\,dx - \int_\Omega |h(0, x)|\,dx \tag{2.9}$$

for any $h \in W^{1,1}(0, T : L^1(\Omega))$, we have that

$$\begin{aligned} &\int_\Omega |S(t, x) - S^*(t, x)|\,dx + \int_0^t \int_\Omega g(C^*)|f(S) - f(S^*)|\,dx\,d\tau \\ &\quad \le \int_\Omega |S_0(x) - S_0^*(x)|\,dx + \int_0^t \int_\Omega f(S)|g(C) - g(C^*)|\,dx\,d\tau, \end{aligned} \tag{2.10}$$

where we have used the fact that $f(S) \ge 0$ if $S \ge 0, g(C) \ge 0$ if $C \ge 0, \text{sign}(S - S^*) = \text{sign}(f(S) - f(S^*))$, and $(f(S) - f(S^*))\text{sign}(f(S) - f(S^*)) = |f(S) - f(S^*)|$. Analogously multiplying (2.8) by $\text{sign}(C - C^*)$ and using the fact that

$$-\int_\Omega \Delta(C - C^*)\text{sign}(C - C^*)\,dx \ge \frac{1}{\alpha}\int_{\partial\Omega} |C - C^*|\,d\sigma, \tag{2.11}$$

we conclude (as before) that

$$\begin{aligned} &\varepsilon \int_\Omega |C(t, x) - C^*(t, x)|\,dx + \frac{1}{\alpha}\int_0^t \int_{\partial\Omega} |C - C^*|\,d\sigma \\ &\qquad + \lambda \int_0^t \int_\Omega f(S)|g(C) - g(C^*)|\,dx\,d\tau \\ &\quad \le \varepsilon \int_\Omega |C_0(x) - C_0^*(x)|\,dx + \lambda \int_0^t \int_\Omega |f(S) - f(S^*)|g(C^*)\,dx\,d\tau. \end{aligned} \tag{2.12}$$

Multiplying (2.10) by λ and adding the result to (2.12) we obtain the conclusion (2.6). Inequalities (2.9) and (2.11) are justified as usual in the L^1 theory of evolution equations by regularizing the sign function. Finally, if f and g are not strictly increasing functions we approximate them by strictly increasing functions and pass to the limit. □

Note that in the case of $\alpha = 0$, the boundary term is absent in (2.6).

For the case of the pseudo-steady-state problem we have the following theorem.

THEOREM 2.3. *Let* f *and* g *be continuous nondecreasing functions with* $f(0) = g(0) = 0$. *Then the problem* $(\hat{\text{P}})$ *has a unique solution* $(\hat{S}, \hat{C})$.

Proof. By the same arguments as in Theorem 2.2 we deduce that if $(\hat{S}, \hat{C})$ and $(\hat{S}^*, \hat{C}^*)$ are two solutions then

$$\begin{aligned} &\lambda \int_\Omega |\hat{S}(t, x) - \hat{S}^*(t, x)|\,dx + \frac{1}{\alpha}\int_0^t \int_{\partial\Omega} |\hat{C}(\tau, \sigma) - \hat{C}^*(\tau, \sigma)|\,d\sigma d\tau \\ &\quad \le \lambda \int_\Omega |\hat{S}(0, x) - \hat{S}^*(0, x)|\,dx. \end{aligned}$$

In particular, as $\hat{S}(0,x) = \hat{S}^*(0,x)$ we conclude that $\hat{S} \equiv \hat{S}^*$ and that $\hat{C}$ and $\hat{C}^*$ are solutions of the elliptic problem

$$-\Delta u + B_t(x,u) = 0 \quad \text{in } \Omega,$$
$$\alpha \frac{\partial u}{\partial \nu} + u = 1 \quad \text{on } \partial\Omega,$$

where $B_t(x,r) = f(\hat{S}(t,x))g(r)$ for any $r \in \mathbb{R}, x \in \Omega$, and $t \in (0,T)$ (here t is a parameter). The uniqueness of $u = \hat{S} = \hat{S}^*$ is now a well-known result since B is monotone nondecreasing in r. □

Remark 2.3. Theorem 2.2 improves a previous result due to Pao [22] where the nonlinearities are assumed to be Lipschitz continuous. Some papers where an L^1 technique is used for parabolic systems are [6], [15], and [33].

Remark 2.4. Basing themselves on our earlier paper [10], DiLiddo and Maddalena were able to prove existence for a different type of problem arising in chemical engineering [12].

3. Asymptotic behavior and monotonicity.

3.1. Monotonicity in λ and initial data. Monotone behavior of the solution of (P′) with respect to initial data and with respect to λ are easy to prove. Monotonicity in time and with respect to ε will require a condition on $C_0(x)$.

PROPERTY I. *For fixed λ and ε, the solutions of* (P′) *are ordered according to their initial values: if $(X_0^{(1)}, C_0^{(1)}) \le (X_0^{(2)}, C_0^{(2)})$ then the respective solutions of* (P′) *satisfy*

$$(X^{(1)}(t,x), C^{(1)}(t,x)) \le (X^{(2)}(t,x), C^{(2)}(t,x)) \quad \textit{for all } x,t.$$

The result follows from the observation that $(X^{(2)}(t,x), C^{(2)}(t,x))$ is a supersolution of problem (P′) with initial data $(X_0^{(1)}, C_0^{(1)})$.

PROPERTY II. *For fixed ε and initial data, the solutions of* (P′) *are ordered inversely with λ:*

$$\lambda_1 \ge \lambda_2 \Rightarrow (X^{(1)}(t,x), C^{(1)}(t,x)) \le (X^{(2)}(t,x), C^{(2)}(t,x)).$$

Again, the proof consists of noting that $(X^{(2)}, C^{(2)})$ is a supersolution of (P′) with $\lambda = \lambda_1$.

Monotonicity with respect to t is a bit more subtle. It is obvious from (2.2) that $X(t,\cdot)$ is monotonically increasing. Since $C_0(x) \le 1$ and the steady state is $C_\infty(x) = 1$, we can only hope to show that $C(t,\cdot)$ is monotonically increasing, but, unfortunately, this cannot be true for all $C_0(x)$. Indeed at $t = 0, C_t \ge 0$ only if

$$-\Delta C_0 + \lambda f(S_0(x)) C_0(x) \le 0. \tag{3.1}$$

Rather than (3.1) we prefer to use the condition

$$-\Delta C_0 + \lambda C_0(x) \le 0, \qquad x \in \Omega, \tag{3.2}$$

which clearly implies (3.1). We also need a condition on the boundary values of $C_0(x)$:

$$C_0 + C_{0,\nu} - 1 \le 0, \qquad x \in \partial\Omega. \tag{3.3}$$

We can then conclude with the following property (see Theorem 3.1).

PROPERTY III. *If $C_0(x)$ satisfies* (3.2) *and* (3.3), *then $t \to (X(t,x), C(t,x))$ is monotone increasing for each x.*

We shall also show (see Lemma 3.1) that under the same conditions on $C_0(x)$ we have monotone behavior with respect to ε. As expected on physical grounds, (X, C) increases as ε decreases.

PROPERTY IV. *For fixed λ and initial data, suppose that $\varepsilon_1 > \varepsilon_2 > 0$ and $C_0(x)$ satisfies* (3.2) *and* (3.3); *then the respective solutions of* (P′) *are ordered so that*

$$(X^1(t,x), C^1(t,x)) \le (X^2(t,x), C^2(t,x)) \quad \textit{for all } (t,x).$$

For the pseudo-steady-state problem only the initial value $\hat{X}_0$ is at our disposal since the initial value $\hat{C}_0$ of the gas concentration is determined from $\hat{X}_0(x)$ as the solution of the elliptic problem

$$\begin{aligned} &-\Delta \hat{C}_0(x) + \lambda F(\hat{X}_0) g(\hat{C}_0) = 0, \qquad x \in \Omega, \\ &\hat{C}_0 + \hat{C}_{0,\nu} = 1, \qquad x \in \partial\Omega. \end{aligned} \tag{3.4}$$

The following results are then easily obtained.

PROPERTY I*. *For fixed λ, the solutions of* ($\hat{\text{P}}'$) *are ordered according to the initial values of $\hat{X}$: if $\hat{X}_0^{(1)} \le \hat{X}_0^{(2)}$, then*

$$\hat{C}_0^{(1)} \le \hat{C}_0^{(2)} \quad \textit{and} \quad (\hat{X}^{(1)}(t,x), \hat{C}^{(1)}(t,x)) \le (\hat{X}^{(2)}(t,x), \hat{C}^{(2)}(t,x)) \quad \textit{for all } (t,x).$$

This follows from the maximum principle or by observing that $\hat{C}_0^{(1)}$ is a subsolution to the scalar elliptic problem for $\hat{C}_0^{(2)}$.

PROPERTY II*. *For fixed initial $\hat{X}_0$, the solutions of* ($\hat{\text{P}}'$) *are ordered inversely with λ:*

$$\lambda_1 \ge \lambda_2 \Rightarrow (\hat{X}^{(1)}, \hat{C}^{(1)}) \le (\hat{X}^{(2)}, \hat{C}^{(2)}) \quad \textit{for all } (t,x).$$

Monotonicity with time is now automatic (see Theorem 3.2).

PROPERTY III*. *$t \to (\hat{X}(t,x), \hat{C}(t,x))$ is monotone increasing for all x.*

In this section we also discuss the behavior as $t \to \infty$. At the simplest level we show that $(X, C) \to (1,1)$ and $(\hat{X}, \hat{C}) \to (1,1)$ as expected. We also discuss the asymptotic limit of (P′) as $\varepsilon \to 0$ and show the various ways in which the solution of (P′) tends to the solution of the pseudo-steady-state problem ($\hat{\text{P}}'$). This relationship requires us to take into account the fact that since $C_0(x) \ne \hat{C}_0(x)$, the limit cannot hold for $t = 0$.

THEOREM 3.1. *Let $0 \le X_0(x) \le 1$, and let $0 \le C_0(x) \le 1$, with $C_0(x)$ satisfying* (3.2) *and* (3.3). *Then the solution (X, C) of* (P′) *has the properties*

$$t \to (X(t,x), C(t,x)) \textit{is monotonically increasing for any } x \in \Omega$$

and

$$\lim_{t\to\infty} (X, C) = (1,1) \quad \textit{in } C([0,\infty) \times \bar{\Omega}).$$

Proof. It is easy to see that $(X_0(x), C_0(x))$ is a subsolution of (P′). We conclude from Theorems 2.1 and 2.2 that

$$X_0(x) \le X(t,x) \le 1, \qquad C_0(x) \le C(t,x) \le 1.$$

From these inequalities, we see at once that, for any $h > 0$, $(X(t+h,x), C(t+h,x))$ is a supersolution of (P′) so that $X(t+h,x) \geq X(t,x)$ and $C(t+h,x) \geq C(t,x)$. Hence (X,C) increases monotonically in time. By the monotone convergence theorem there exists $(X_\infty(x), C_\infty(x))$ with $0 \leq X_\infty \leq 1, 0 \leq C_\infty \leq 1$ such that $\lim_{t\to\infty}(X,C) = (X_\infty, C_\infty)$ in $L^p(\Omega)$ for any p with $1 \leq p \leq \infty$. On the other hand, using the definition of weak solutions and the monotonicity of F and g, it is not difficult to show (see the argument in Sattinger [23]) that (X_∞, C_∞) must be the solution of the stationary problem. □

COROLLARY 3.1. *If* $0 \leq X_0(x) \leq 1$ *and* $0 \leq C_0(x) \leq 1$, *then the solution of* (P′) *satisfies* $\lim_{t\to\infty}(X,C) = (1,1)$.

Proof. Since C_0 does not necessarily satisfy (3.2) and (3.3), $C(t,\cdot)$ may not be monotone. Consider, however, the solution $(X\#, C\#)$ of (P′) with initial data $(X_0, 0)$. Then $(X\#, C\#)$ is easily seen to be a subsolution of (P′) with initial data (X_0, C_0) so that

$$0 \leq X\# \leq X \leq 1, \qquad 0 \leq C\# \leq C \leq 1.$$

By Theorem 3.1, $(X\#, C\#)$ tends monotonically to $(1,1)$ as $t \to \infty$, so that (X,C) also tends to $(1,1)$ as $t \to \infty$ (but perhaps not monotonically). □

For the pseudo-steady-state problem $(\hat{\text{P}})$, the monotonicity in time of $\hat{X}$ and $\hat{C}$ is always guaranteed. The straightforward proof is omitted.

THEOREM 3.2. *If* $(\hat{X}, \hat{C})$ *is the solution of* $(\hat{\text{P}}')$, *then*

$$t \to (\hat{X}(t,x), \hat{C}(t,x)) \text{ is monotonically increasing}$$

for any $x \in \Omega$ *and* $\lim_{t\to\infty}(\hat{X}, \hat{C}) = (1,1)$.

3.2. Monotonicity in ε and the relationship between (P) and $(\hat{\text{P}})$. Consider problem (P′) for fixed λ and fixed initial data (X_0, C_0), but with different values of ε. To emphasize the dependence on ε we relabel the problem (P′) as (P'_ε) and its solution as $(X^\varepsilon, C^\varepsilon)$. When is there monotonicity of (P'_ε) with respect to ε? In practice, ε is often small and problem $(\hat{\text{P}}')$ with initial value $\hat{X}_0 = X_0$ (and $\hat{C}_0$ determined from (3.4)) is used to approximate (P'_ε). In what sense, if any, is this a good approximation?

We begin with two simple lemmas.

LEMMA 3.1. *Let* $\varepsilon_1 \geq \varepsilon_2 \geq 0$; *let* $C_0(x)$ *satisfy* (3.2) *and* (3.3); *and let* (X^i, C^i), $i = 1,2$ *be the solutions of* (P'_ε) *corresponding to* $\varepsilon = \varepsilon_i$ *and initial data independent of* i. *Then*

$$(X^1, C^1) \leq (X^2, C^2) \quad \text{for any } (t,x).$$

Proof. Since $X^1_t \geq 0$ and, by Theorem 3.1, $C^1_t \geq 0$, we have $\varepsilon_2 C^1_t - \Delta C^1 + \lambda F(X^1) g(C^1) = (\varepsilon_2 - \varepsilon_1) C^1_t \leq 0$, so that (X^1, C^1) is a lower solution to (P'_ε) with $\varepsilon = \varepsilon_2$. Hence, for all (t,x), we have

$$(X^1(t,x), C^1(t,x)) \leq (X^2(t,x), C^2(t,x)) \leq (1,1). \qquad \square$$

Remark 3.1. If, for instance, (3.2) does not hold at some point x, then we can have $C^1(t,x) > C^2(t,x)$ for small t.

LEMMA 3.2. *If* $(X_0, C_0) \leq (\hat{X}_0, \hat{C}_0)$ *then* $(X(t,x), C(t,x)) \leq (\hat{X}(t,x)\hat{C}(t,x))$.

Proof. We have

$$\varepsilon \hat{C}_t - \Delta \hat{C} + \lambda F(\hat{X}) g(\hat{C}) = \varepsilon \hat{C}_t \geq 0,$$

where the last inequality follows from Theorem 3.2. In view of the assumption on the initial values, $(\hat{X}, \hat{C})$ is a supersolution of (P′) and the result follows. □

Remark 3.2. If $X_0 = \hat{X}_0$ and $C_0 = \hat{C}_0$, we conclude that problem $(\hat{\mathrm{P}}')$ converts solid more quickly than (P'), as expected.

These two lemmas lead to the following theorem, which deals with the limit as $\varepsilon \to 0$ of (P'_ε).

THEOREM 3.3. *If* $(X_0, C_0) \leq (\hat{X}_0, \hat{C}_0)$, *then* $(X^\varepsilon, C^\varepsilon) \leq (\hat{X}, \hat{C})$ *and* $\lim_{\varepsilon \to 0}(X^\varepsilon, C^\varepsilon) = (\hat{X}, \hat{C})$ *in* $C((0,\infty) \times \Omega)$. *Moreover* $\varepsilon \to X^\varepsilon(t,x)$ *is monotone increasing for any* t, x, *and if* $C_0(x)$ *also satisfies* (3.2) *and* (3.3), $\varepsilon \to C^\varepsilon(t,x)$ *is monotone increasing as well.*

Proof. Suppose C_0 satisfies (3.2), (3.3); then, by Lemma 3.1, $(X^\varepsilon, C^\varepsilon)$ is monotone in ε and so converges to $(\tilde{X}, \tilde{C})$ as $\varepsilon \to 0+$, uniformly on $\bar{\Omega}_T$; moreover, $(\tilde{X}, \tilde{C})$ clearly satisfies the pseudo-steady-state problem $(\hat{\mathrm{P}}')$ and therefore $\tilde{X} = \hat{X}, \tilde{C} = \hat{C}$. Now let $(X^\varepsilon_\#, C^\varepsilon_\#)$ be the unique solution of $(\mathrm{P}'_\varepsilon)$ with initial data $(X_0, 0)$; then $(X^\varepsilon_\#, C^\varepsilon_\#)$ is seen to be a subsolution to $(\mathrm{P}'_\varepsilon)$ for the same ε and initial data (X_0, C_0). It then follows from Lemma 3.2 that

$$(X^\varepsilon_\#, C^\varepsilon_\#) \leq (X^\varepsilon, C^\varepsilon) \leq (\hat{X}, \hat{C}).$$

But $(X^\varepsilon_\#, C^\varepsilon_\#)$ satisfies the conditions of Lemma 3.1, so it is monotone increasing in ε and therefore tends to $(\hat{X}, \hat{C})$ as $\varepsilon \to 0+$. Hence, so does $(X^\varepsilon, C^\varepsilon)$. □

In the following theorem we provide other measures of how well $(\hat{\mathrm{P}}')$ approximates $(\mathrm{P}'_\varepsilon)$ for ε small. The initial data for $(\mathrm{P}'_\varepsilon)$ is (X_0, C_0), and for $(\hat{\mathrm{P}}')$ is $(\hat{X}_0 = X_0, \hat{C}_0)$, where $\hat{C}_0$ is determined from (3.4).

THEOREM 3.4. *Let* $X_0 \in C^\delta(\Omega)$ *for some* $\delta \in (0,1)$ *and let* $C_0 \in C^{2+\delta}(\Omega)$ *with* $(0,0) \leq (X_0, C_0) \leq (1,1)$. *Then the estimate* (2.6) *holds replacing* $(S^* = 1 - X^*, C^*)$ *by* $(\hat{S} = 1 - \hat{X}, \hat{C})$. *In particular,*

$$\|X^\varepsilon(t,x) - \hat{X}(t,x)\|_{L^1(\Omega)} \leq M\varepsilon \tag{3.5}$$

for any $t \geq 0$ (*i.e.,* $X^\varepsilon \to \hat{X}$ *in* $C([0,\infty) : L^1(\Omega))$ *as* $\varepsilon \to 0$), *where* $M = \frac{1}{\lambda}\|C_0 - \hat{C}_0\|_{L^1(\Omega)}$. *Moreover, if*

$$\|C_0 - \hat{C}_0\|_{L^1(\Omega)} \leq L\varepsilon^\gamma \quad \text{for some } L > 0 \text{ and } \gamma > 0, \tag{3.6}$$

then

$$\|X^\varepsilon(t,x) - \hat{X}(t,x)\|_{L^1(\Omega)} \leq \varepsilon^{\gamma+1}\frac{L}{\lambda} \tag{3.7}$$

and

$$\|C^\varepsilon(t,x) - \hat{C}(t,x)\|_{L^1(\Omega)} \leq L\varepsilon^\gamma \tag{3.8}$$

for any $t \geq 0$ (*i.e.,* $(X^\varepsilon, C^\varepsilon) \to (\hat{X}, \hat{C})$ *in* $C([0,\infty) : L^1(\Omega))$ *as* $\varepsilon \to 0$).

Proof. By Theorem 3.2 we have $\hat{C}_t \geq 0$ so that $\varepsilon\hat{C}_t - \Delta\hat{C} + \lambda F(\hat{X})g(\hat{C}) \geq 0$ and the proof of Theorem 2.2 gives the inequality

$$\varepsilon \int_\Omega |C^\varepsilon(t,x) - \hat{C}(t,x)|\,dx + \lambda \int_\Omega |X^\varepsilon(t,x) - \hat{X}(t,x)|\,dx + \frac{1}{\alpha}\int_0^t \int_{\partial\Omega} |C^\varepsilon(\tau,\sigma) - \hat{C}(\tau,\sigma)|\,d\tau\,d\sigma \leq \varepsilon \int_\Omega |C_0(x) - \hat{C}_0(x)|\,dx,$$

which leads to the desired results (3.7) and (3.8). □

The next result shows the convergence of $(X^\varepsilon, C^\varepsilon)$ to $(\hat{X}, \hat{C})$ as $\varepsilon \to 0$ in the space $L^1(0,t : C(\bar{\Omega}))$ independently of the initial difference $\|C_0 - \hat{C}_0\|_{L^\infty(\Omega)}$.

THEOREM 3.5. *Let X_0, C_0 be as in Theorem* 3.4 *and let $C_0 \le \hat{C}_0$. Let $w \in C^\infty(\Omega)$ be the (unique) solution of the linear problem* (1.11). *Then, for any $t > 0$ and any $x \in \bar{\Omega}$,*

$$0 \le \int_0^t (\hat{C}(\tau, x) - C^\varepsilon(\tau, x))\, d\tau \le \varepsilon w(x) \tag{3.9}$$

and

$$0 \le \int_\Omega (\hat{X}(t, x) - X^\varepsilon(t, x))\, dx \le \varepsilon \frac{|\Omega|}{\lambda}, \tag{3.10}$$

so that

$$C^\varepsilon \to \hat{C} \quad \text{in } L^1(0,t : C(\bar{\Omega}))$$

and

$$X^\varepsilon \to \hat{X} \quad \text{in } L^\infty(0,t : L^1(\bar{\Omega})).$$

Proof. Integrate the equations for C^ε and $\hat{C}$ with respect to time; use Lemma 3.2 and the comparison principle to obtain (3.9). Inequality (3.10) follows from Green's formula.

4. On the conversion of the solid and the penetration of the gas.[1]

4.1. Solid conversion. The behavior of $f(S)$ near $S = 0$ will determine whether or not the solid is fully converted in finite time.

THEOREM 4.1. *Consider problems* (P) *and* ($\hat{\text{P}}$) *and let*

$$R(S) = \int_S^1 \frac{d\sigma}{f(\sigma)}, \qquad I = R(0+). \tag{4.1}$$

Then if $I < \infty$, the solid is fully converted in finite time; if $I = \infty, S(x,t)$ and $\hat{S}(x,t)$ are positive for all t at every x where the initial solid concentration is positive.

Proof. We carry out the proof for problem (P), the reasoning being similar for ($\hat{\text{P}}$). We write (1.1) with its initial condition (1.4a) as

$$S_t = -\mu(t,x) f(S), \quad t > 0, \qquad S(0,x) = S_0(x), \tag{4.2}$$

and treat each x individually. If $S_0(\xi) = 0$, then $S(t,\xi) \equiv 0$ for all t, so we confine ourselves to points x where $S_0(x) > 0$. In that case, $S(t,x)$ will itself be positive for some initial time interval, and we can divide (4.2) by $f(S)$ to obtain

$$\frac{d}{dt} R(S) = \mu(t,x), \tag{4.3}$$

where $R(S)$ is defined by (4.1). Note that $R(S)$ is positive and decreasing on $0 < S < 1$. Therefore, $R^{-1}(S)$ is positive and decreasing on $[0, I)$. Integrating (4.3) from 0 to t gives

$$R(S) = R(S_0) + \int_0^t \mu(\tau, x)\, d\tau,$$

[1] In this section $\|\cdot\|$ will stand for the usual sup norm on the x variable, $\|u(t,x)\| = \sup_{x \in \Omega} \|u(t,x)\|$.

and therefore

$$S(t,x) = R^{-1}\left[\int_0^t \mu(\tau,x)\,d\tau + R(S_0(x))\right] \tag{4.4}$$

is the solution of (4.2) as long as $S > 0$. If $I = \infty$, the positivity of R^{-1} on $[0,\infty)$ guarantees that $S(t,x) > 0$, so that the second part of the theorem is proved. If I is finite, (4.4) furnishes a positive solution of (4.2) for $t < T(x)$ where $T(x)$ is defined by

$$R(S_0(x)) + \int_0^T \mu(\tau,x)\,d\tau = I.$$

At $t = T(x), S(x,t) = 0$ and remains equal to zero for $t \geq T(x)$. Since $R^{-1}(I) = 0$, it is useful to extend the definition of R^{-1} through the rule $R^{-1}(z) = 0, z \geq I$. With that agreement (4.4) remains valid for all t even if I is finite.

Returning to (1.1) we can then write

$$\begin{aligned} S(t,x) &= R^{-1}\left[\int_0^t g(C(\tau,x))\,d\tau + R(S_0(x))\right] \\ &\leq R^{-1}\left[\int_0^t g(C(\tau,x))\,d\tau\right]. \end{aligned} \tag{4.5}$$

We have proved in §3 that $C(t,x)$ tends to 1 as $t \to \infty$, uniformly for $x \in \bar{\Omega}$, so that $\int_0^t g(C(\tau,x))\,d\tau \geq I$ for all x if t is sufficiently large. Therefore, there exists T such that $S(t,x) \equiv 0, t \geq T$, and we have *full conversion in finite time.* □

COROLLARY 4.1. *If $f(S) = S^m, I$ is finite if and only if $m < 1$. The explicit formulas are*

$$R_m(S) = \begin{cases} \frac{1-S^{1-m}}{1-m}, & m \neq 1, \\ -\ln S, & m = 1, \end{cases}$$

and thus

$$R_m^{-1}(z) = \begin{cases} [1 - z(1-m)]_+^{1/1-m}, & m \neq 1, \\ e^{-z}, & m = 1, \end{cases} \tag{4.6}$$

where $[u]_+$ stands for the greater of u and 0. We can then rewrite (4.5) as

$$S(t,x) = S_0(x) R_m^{-1}\left[S_0^{-1}(x)\int_0^t g(C(\tau,x))\,d\tau\right].$$

Remark 4.1. We could substitute (4.5) into (1.2) to obtain a nonlinear integro-differential equation for C subject to conditions (1.4b) and (1.4c). A related approach due to McNabb [19] is more useful. He introduces

$$\eta(t,x) = \int_0^t [1 - C(\tau,x)]\,d\tau \qquad (1 - C = \eta_t), \tag{4.7}$$

the time-integrated deviation of C from its steady state. A straightforward calculation shows that η satisfies

$$\begin{aligned} &\varepsilon\eta_t - \Delta\eta = \varepsilon(1 - C_0) + \lambda(S_0 - S), \quad x \in \Omega, \quad t > 0; \\ &\eta(x,0) = 0, \quad \eta + \alpha\eta_\nu = 0, \quad x \in \partial\Omega, \quad t > 0. \end{aligned} \tag{4.8}$$

Although S is unknown in (4.8), considerable information can nevertheless be extracted from this formulation. For instance, because S decreases to 0 as $t \to \infty, \eta(t, \cdot)$ is monotonically increasing to the solution $\eta_\infty(x)$ of the steady-state problem

$$-\Delta\eta_\infty = \varepsilon(1 - C_0) + \lambda S_0, \quad \eta_\infty + \alpha\eta_{\infty,\nu} = 0, \quad x \in \partial\Omega. \tag{4.9}$$

Note that

$$\eta_\infty \leq (\varepsilon + \lambda)w(x), \tag{4.10}$$

where $w(x)$ is the solution of (1.11).

For the p.s.s. problem we define

$$\hat{\eta} = \int_0^t (1 - \hat{C})\, d\tau, \tag{4.11}$$

which satisfies

$$-\Delta\hat{\eta} = \lambda(\hat{S}_0 - \hat{S}), \quad \hat{\eta} + \alpha\hat{\eta}_\nu = 0, \quad x \in \partial\Omega, \tag{4.12}$$

and $\hat{\eta}(t, x)$ tends monotonically to the solution of

$$-\Delta\hat{\eta}_\infty = \lambda\hat{S}_0, \quad \hat{\eta}_\infty + \alpha\hat{\eta}_{\infty,\nu} = 0, \quad x \in \partial\Omega. \tag{4.13}$$

We see that

$$\hat{\eta}_\infty \leq \lambda w(x). \tag{4.14}$$

In the *special case where* $g(C) = C$, (4.5) gives

$$S(t, x) = R^{-1}[t - \eta + R(S_0)],$$

which can be substituted into (4.8) to give a scalar partial differential equation for η. These ideas were exploited in [24] and [28] and will be used to some extent in the remainder of the section.

Remark 4.2. If we had considered a problem *without gas diffusion* and with a gas concentration *maintained at the value one,* the solid concentration $S^*(t, x)$ would satisfy the ordinary differential equation

$$S_t^* = -f(S^*), \quad t > 0, \quad S^*(0, x) = S_0^*(x). \tag{4.15}$$

There are many ways of seeing that $S^*(t, x) \leq S(t, x)$, where S is the solution of (P) with the same initial solid concentration (and any $C_0 \leq 1$). For instance, since R^{-1} is monotone decreasing and $g(C) \leq 1$, (4.5) shows that

$$S(t, x) \geq R^{-1}(t + R(S_0(x))),$$

the right-hand side being precisely $S^*(t, x)$, because now $\mu \equiv 1$ in (4.4). Similarly, we can show $S^* \leq \hat{S}$, the solution of ($\hat{\text{P}}$) with the same initial value.

The quantity that is perhaps of greatest physical interest is the overall conversion fraction $\gamma(t)$. The inverse of this function gives the time required to achieve the

conversion of a specified fraction of the solid. We shall compare problems (P), (P′), and (4.15) with the same initial value $S_0(x)$.

The *overall conversion* at time t for problem (P) is given by

$$\gamma(t) = \frac{\int_\Omega (S_0(x) - S(t,x))\,dx}{\int_\Omega S_0(x)\,dx} = 1 - \frac{\int_\Omega S(t,x)\,dx}{\int_\Omega S_0(x)\,dx}, \tag{4.16}$$

for problem ($\hat{\text{P}}$) by

$$\hat{\gamma}(t) = 1 - \frac{\int_\Omega \hat{S}(t,x)\,dx}{\int_\Omega S_0(x)\,dx},$$

and for (4.15) by

$$\gamma^*(t) = 1 - \frac{\int_\Omega S^*(t,x)\,dx}{\int_\Omega S_0(x)\,dx}.$$

We have already proved the following properties:

(a) $0 \le \gamma(t) \le 1, \quad 0 \le \hat{\gamma}(t) \le 1, \quad 0 \le \gamma^*(t) \le 1;$

(b) $\gamma(t) \le \hat{\gamma}(t), \quad \hat{\gamma}(t) = \gamma^*(t)$ (see Remark 4.2);

(c) *If* $C_0(x) \le \hat{C}_0(x), \quad \gamma(t) \le \hat{\gamma}(t)$ (see Lemma 3.2);

(d) $\lim_{t\to\infty} \gamma(t) = \lim_{t\to\infty} \hat{\gamma}(t) = \lim_{t\to\infty} \gamma^*(t) = 1$ (see Corollary 3.1).

If I is finite, Theorem 4.1 tells us that the solid is fully converted in finite time, that is, $\gamma(t) \equiv 1$ for t sufficiently large. We define full conversion times $t_1, \hat{t}_1, t_1^*$ by

$$t_1 = \inf\{t : \gamma(t) = 1\}, \quad \hat{t}_1 = \inf\{t : \hat{\gamma}(t) = 1\}, \quad t_1^* = \inf\{t : \gamma^*(t) = 1\}.$$

We first observe that t_1^* is known explicitly. We seek the smallest value of t for which $S^*(t,x)$, the solution of (4.15), is identically zero on $\bar{\Omega}$. Since $\|S_0\| = 1$, there is at least one point ξ where $S_0(\xi) = 1$. These points will be the slowest to convert. From (4.4) we have $S(t,\xi) = R^{-1}(t)$, which is positive for $t < I$ and vanishes for $t \ge I$. Therefore, $t_1^* = I$.

From (b) and (c) above we see that

$$t_1 \ge I, \quad \hat{t}_1 \ge I \quad \text{and, if} \quad C_0(x) \le \hat{C}_0(x), \quad t_1 \ge \hat{t}_1.$$

We can obviously characterize t_1 by

$$\inf\{t : S(t,x) \equiv 0, x \in \bar{\Omega}\} = \inf\{t : X(t,x) \equiv 1, x \in \bar{\Omega}\}.$$

By (4.5) and the fact that $R^{-1}(I) = 0$, we also have that t_1 is characterized by

$$\min_{x\in\bar{\Omega}} \int_0^{t_1} g(C(\tau,x))\,d\tau + R(S_0(x)) = I$$

and $\hat{t}_1$ by

$$\min_{x\in\bar{\Omega}} \int_0^{\hat{t}_1} g(\hat{C})\,d\tau + R(\hat{S}_0) = I.$$

If $S_0(x) = \hat{S}_0(x) \equiv 1$, then we see that t_1 and $\hat{t}_1$ satisfy

$$\min_{x\in\bar{\Omega}} \int_0^{t_1} g(C)\,d\tau = I, \qquad \min_{x\in\bar{\Omega}} \int_0^{\hat{t}_1} g(\hat{C})\,d\tau = I. \tag{4.17}$$

Next, we provide estimates for $\hat{t}_1$ when $\hat{S}_0(x) \equiv 1$, the case that occurs most frequently in applications (uniform initial solid concentration).

THEOREM 4.2. *Let $\hat{S}_0 \equiv 1$ and let I be finite. Assume there exist $g_1, g_2 \in C^0([0,1]) \cap C^1((0,1))$ such that $g_1', g_2' \geq 0, g_1(1) = g_2(1) = 1$, and*

$$g_1(r) \leq g(r) \leq g_2(r) \quad \forall\, r \in [0,1]. \tag{4.18}$$

Then

$$I + M_2\lambda\|w\| \leq \hat{t}_1 \leq I + M_1\lambda\|w\|, \tag{4.19}$$

where $M_1 = \sup_{[0,1]} g_1'$, $M_2 = \inf_{[0,1]} g_2'$, and w is defined by (3.9).

Proof. By definition, $\hat{\eta}_t = 1 - \hat{C}$ so that $0 \leq 1 - \hat{\eta}_t \leq 1, 0 \leq \hat{\eta}_t \leq 1$. We immediately see that with M_1, M_2 as defined above,

$$1 - g_1(1 - \hat{\eta}_t) \leq M_1\hat{\eta}_t, \qquad 1 - g_2(1 - \hat{\eta}_t) \geq M_2\hat{\eta}_t,$$

and, therefore,

$$1 - M_1\hat{\eta}_t \leq g_1(1 - \hat{\eta}_t) \leq g(1 - \hat{\eta}_t) \leq g_2(1 - \hat{\eta}_t) \leq 1 - M_2\hat{\eta}_t.$$

Integrating from 0 to $\hat{t}_1$, we find

$$t_1 - M_1\hat{\eta}(\hat{t}_1, x) \leq \int_0^{\hat{t}_1} g(1 - \hat{\eta}_t)\, d\tau \leq \hat{t}_1 - M_2\hat{\eta}(\hat{t}_1, x),$$

and, taking the minimum with respect to x and using the characterization (4.17) for $\hat{t}_1$, we obtain

$$\hat{t}_1 - M_1\|\hat{\eta}(\hat{t}_1, x)\| \leq I \leq \hat{t}_1 - M_2\|\hat{\eta}(\hat{t}_1, x)\|. \tag{4.20}$$

When $t \geq \hat{t}_1, \hat{S} = 0$ so that (4.12) gives

$$\hat{\eta} = \hat{\eta}_\infty = \lambda w(x), \qquad \|\hat{\eta}(t_1, x)\| = \lambda\|w\|,$$

which, when substituted into (4.20) gives the result (4.19). □

Remark 4.3. Special cases of interest correspond to $g(C) = C^p, p > 0$. Our estimates then become

(4.21) (a) if $p = 1$, $\hat{t}_1 = I + \lambda\|w\|$ (take $g_1 = g_2 = g$);

(b) if $p < 1$,

(4.22) $I + p\lambda\|w\| \leq \hat{t}_1 \leq I + \lambda\|w\|$ (take $g_1(r) = r, g_2 = g$);

(c) if $p > 1$,

(4.23) $I + \lambda\|w\| \leq \hat{t}_1 \leq I + \lambda p\|w\|$ (take $g_2(r) = r, g_1 = g$).

The result (a) was first given in [28]. From the inequality $0 \leq \int_0^{\hat{t}_1}(1 - \hat{\eta}_t)\, d\tau = \hat{t}_1 - \lambda w(x)$, we also find

$$\hat{t}_1 \geq \lambda\|w\|, \tag{4.24}$$

which improves the lower bound (b) if $\lambda(1-p)\|w\| \geq I$.

Further improvement follows from using Jensen's inequality, which we illustrate in the case $p < 1$, when the inequality becomes

$$\left(\int_0^t v(\tau)\,d\tau\right)^p \geq t^{p-1}\int_0^t v^p(\tau)\,d\tau,$$

where $v(t)$ is any nonnegative continuous function.

Setting $v = 1 - \hat{\eta}_t$, we find

$$\int_0^{\hat{t}_1} (1-\hat{\eta}_t)^p\,d\tau \leq \hat{t}_1^{1-p}(\hat{t}_1 - \lambda w(x))^p \quad \forall x \in \bar{\Omega},$$

and, taking the minimum over x,

$$I \leq \hat{t}_1^{1-p}(\hat{t}_1 - \lambda\|w\|)^p.$$

This inequality implies $\hat{t}_1 \geq T$, where T is the unique positive solution of

$$I = T^{1-p}(T - \lambda\|w\|)^p, \tag{4.25}$$

which can be shown to always provide a better lower bound to $\hat{t}_1$ than (4.22) and (4.24). As an example, if $p = \frac{1}{2}$, (4.25) gives

$$T = \frac{\lambda\|w\| + \sqrt{\lambda^2\|w\|^2 + I^2}}{2},$$

which is larger than the lower bounds (4.22) and (4.24).

We now turn to estimates for t_1 when $S_0(x) \equiv 1$. These are somewhat more difficult as t_1 also depends on both ε and $C_0(x)$. We will not be able, for instance, to find an exact value for t_1 when $p = 1$, as we did for $\hat{t}_1$ (see (4.21)).

THEOREM 4.3. *Let* $S_0(x) \equiv 1, 0 \leq C_0(x) \leq 1$; *then*

$$t_1 \leq I + M_1(\varepsilon + \lambda)\|w\|,$$

and if also $C_0 \leq \hat{C}_0$, *then*

$$t_1 \geq \hat{t}_1 \geq I + M_2\lambda\|w\|,$$

where M_1, M_2 *are as in Theorem* 4.2.

Proof. We proceed as in Theorem 4.2 to reach the equivalent of (4.20):

$$t_1 - M_1\|\eta(t_1,x)\| \leq I \leq t_1 - M_2\|\eta(t_1,x)\|. \tag{4.26}$$

Unfortunately, $\eta(t_1,x)$ is not known explicitly, so we must use estimates for $\|\eta(t_1,x)\|$. From (4.9) we have

$$\eta(t,x) \leq \eta_\infty(x) \leq (\varepsilon+\lambda)w(x),$$

which yields

$$t_1 \leq I + M_1(\varepsilon+\lambda)\|w\|,$$

the upper bound in Theorem 4.3. The lower bound in the theorem follows from previous results (Theorem 4.2 and Lemma 3.2). □

Remark 4.4. Again we can consider the case C^p and find corresponding forms of (4.21)–(4.24). For instance, (4.21) becomes

$$I + \lambda\|w\| \leq t_1 \leq I + (\lambda + \varepsilon)\|w\|,$$

which is quite satisfactory as long as ε is small.

Remark 4.5. Jensen's inequality can be used to improve some of the bounds.

Remark 4.6. We can also be more accurate in our estimate for $\|\eta(t_1, x)\|$, which appears in the proof of Theorem 4.3. For $t \geq t_1, \eta(t, x)$ satisfies (4.8) with $S_0 \equiv 1, S \equiv 0$; that is,

$$-\Delta\eta = \varepsilon(1 - C_0) + \lambda - \varepsilon\eta_t; \quad \eta + \alpha\eta_\nu = 0 \quad \text{on } \partial\Omega.$$

Now $0 \leq \eta_t \leq 1$, so that

$$\lambda - \varepsilon C_0 \leq -\Delta\eta \leq \varepsilon(1 - C_0) + \lambda$$

and, hence,

$$\lambda w(x) - \varepsilon y(x) \leq \eta(t, x)) \leq (\lambda + \varepsilon)w(x) - \varepsilon y(x),$$

where $y(x)$ satisfies

$$-\Delta y = C_0(x), \quad x \in \Omega; \quad y + \alpha y_\nu = 0 \quad \text{on } \partial\Omega.$$

As an illustration of the use of these inequalities, suppose $C_0 \equiv 1$. Then (4.26) yields

$$t_1 - M_1\lambda\|w\| \leq I \leq t_1 - M_2(\lambda - \varepsilon)\|w\|$$

and

$$I + M_2(\lambda - \varepsilon)\|w\| \leq t_1 \leq I + \lambda M_1\|w\|.$$

If, also, $p = 1$, then with $g_1 = g_2 = g$, we find $M_1 = M_2 = 1$ and

$$I + (\lambda - \varepsilon)\|w\| \leq t_1 \leq I + \lambda\|w\| = \hat{t}_1.$$

Thus, unlike the case where $C_0 \leq \hat{C}_0$, we now have $t_1 \leq \hat{t}_1$.

4.2. Penetration of the gas. In problems (P) and ($\hat{\text{P}}$), the gas concentration tends uniformly to its steady state $C = \hat{C} = 1$. The concentration must therefore be strictly positive for t sufficiently large, but is this necessarily true for all $t > 0$? Our experience with scalar problems involving strong absorption suggests otherwise: there may exist a time-dependent "dead core" in which the concentration is zero at time t (see [4] and [9]). Any such dead core must, of course, disappear in finite time. For problems (P) and ($\hat{\text{P}}$) we define

$$D(t) = \{x \in \Omega : C(t, x) = 0\}, \qquad T = \inf\{t : C(t, x) > 0 \text{ for all } x \in \bar{\Omega}\}, \tag{4.27}$$

$$\hat{D}(t) = \{x \in \Omega : \hat{C}(t, x) = 0\}, \qquad \hat{T} = \inf\{t : \hat{C}(t, x) > 0 \text{ for all } x \in \bar{\Omega}\}. \tag{4.28}$$

As in the scalar case, an important role is played by

$$J = \int_{0+}^{1} \frac{d\sigma}{\sqrt{G(\sigma)}}, \quad \text{where } G(\sigma) = \int_0^\sigma g(C)\, dC. \tag{4.29}$$

Note that if $g(C) = C^p$, J is finite for $0 < p < 1$ and infinite if $p \geq 1$.

Our main result is that (a) no dead core exists if $J = \infty$, and (b) a dead core may exist if J is finite. We prove these assertions below and also provide rough estimates for the location of the dead core and for the time at which it disappears.

THEOREM 4.4. *Let $g(C)$ be such that $J = \infty$. Then*

$$\hat{C}(t,x) > 0 \quad \textit{for all } (t,x) \in [0,\infty) \times \bar{\Omega},$$
$$C(t,x) > 0 \quad \textit{for all } (t,x) \in (0,\infty) \times \bar{\Omega}.$$

Proof. By Theorem 3.2, $\hat{C}(t,x)$ increases monotonically in time, so $\hat{C}(t,x) \geq \hat{C}_0(x)$ with $\hat{C}_0$ defined by (1.9). Thus $\hat{C}_0(x)$ satisfies

$$-\Delta \hat{C}_0 + \lambda g(\hat{C}_0) \geq 0$$

and, therefore, by a result of Vázquez [30], $\hat{C}_0(x) > 0$ in $\bar{\Omega}$, and hence $\hat{C}(t,x) > 0$ in $\bar{\Omega}$ for all $t \geq 0$. In the parabolic case, we have

$$0 = \varepsilon C_t - \Delta C + \lambda f(S) g(C) \leq \varepsilon C_t - \Delta C + \lambda g(C),$$

so that $C(t,x)$ is a supersolution of the scalar problem

$$\varepsilon C^*_t - \Delta C^* + \lambda g(C^*) = 0 \quad \text{on } (0,\infty) \times \Omega,$$
$$C^* + \alpha C^*_\nu = 1 \quad \text{on } (0,\infty) \times \partial\Omega,$$
$$C^*(0,x) = C_0(x) \quad \text{on } \Omega.$$

The strict positivity can be obtained by an easy modification of a result of Bertsch, Kersner, and Peletier [5] and, therefore, $C(t,x) > 0$ on $(0,\infty) \times \bar{\Omega}$. □

Next we show that if J is finite, a dead core is possible in the pseudo-steady-state case.

THEOREM 4.5. *Let $J < \infty$; define*

$$d(x,\partial\Omega) = \textit{ distance from } x \textit{ to } \partial\Omega,$$
$$A = \textit{ half-width of thinnest slab enclosing } \Omega,$$
$$r_i = \textit{ radius of largest inscribed ball,}$$
$$m(t) = f(\inf_{x \in \bar{\Omega}} \hat{S}(t,x)), \qquad M(t) = f(\sup_{x \in \bar{\Omega}} \hat{S}(t,x)).$$

Then

$$\hat{D}(t) \supset \left\{ x \in \Omega : d(x,\partial\Omega) \geq \left[\frac{N}{2\lambda m(t)}\right]^{1/2} J \right\}, \tag{4.30}$$

$$\hat{D}(t) \neq \emptyset \quad \textit{when } \left[\frac{N}{2\lambda m(t)}\right]^{1/2} J < r_i, \tag{4.31}$$

$$\hat{D}(t) = \emptyset \quad \textit{when } \lambda < \frac{J^2}{2a^2 M(t)}. \tag{4.32}$$

$$\hat{D}(t) = \emptyset \quad \textit{for all } t \textit{ if } \lambda < J^2/2a^2. \tag{4.33}$$

Proof. Since f is increasing, we have

$$-\Delta \hat{C} + \lambda M(t) g(\hat{C}) \geq 0 = -\Delta \hat{C} + \lambda f(\hat{S}) g(\hat{C}) \geq -\Delta \hat{C} + \lambda m(t) g(\hat{C}),$$

so that $\hat{C}$ is a subsolution of the scalar elliptic problem

$$-\Delta C^{(\alpha)} + \lambda m(t) g(C^{(\alpha)}) = 0, \qquad x \in \Omega,$$
$$C^{(\alpha)} + \alpha C_\nu^{(\alpha)} = 1, \qquad x \in \partial\Omega.$$

Therefore, $\hat{C}(t,x) \le C^{(\alpha)}(t,x)$, which, in turn, is smaller than the solution $C^{(0)}(t,x)$ of the Dirichlet problem. It follows that $\hat{D}(t)$ is contained in the dead core of $C^{(0)}(t,x)$. It is shown in Diaz [7, Prop. 1.11] that the dead core for $C^{(0)}$ satisfies (4.30); see also [8] and [25]. The assertion (4.31) is an immediate consequence of (4.30).

To prove (4.32) we observe that $\hat{C}$ is a supersolution to the scalar problem with $M(t)$ replacing $f(\hat{S})$. Although conditions for nonexistence of a dead core for this latter problem are available [17], [26], we confine ourselves to the case $\alpha = 0$, when the simple bound (4.32) is derived in [3]. Since $M(t)$ is a decreasing function of time with $M(0) = 1$, the bound (4.33) follows. □

Remark 4.7. To apply (4.31), we need an explicit lower bound for $m(t)$. Such a bound is easily obtained if $\inf_{\bar{\Omega}} \hat{S}_0(x) = \delta > 0$. Then, from Remark 4.2, we find $\hat{S}(t,x) \ge R^{-1}(t + R(\delta))$ so that $m(t) \ge f[R^{-1}(t + R(\delta))]$ and $m(0) \ge f(\delta)$. Therefore, a dead core exists at time t if

$$r_i > J\left\{\frac{N}{2\lambda f[R^{-1}(t + R(\delta))]}\right\}^{1/2}. \tag{4.34}$$

In particular, a dead core exists for sufficiently small t if

$$r_i > J\left\{\frac{N}{2\lambda f(\sigma)}\right\}^{1/2}. \tag{4.35}$$

Note that (4.34) also gives a lower bound for $\hat{T}$:

$$\hat{T} \ge R\left\{f^{-1}\left(\frac{J^2 N}{2\lambda r_i^2}\right) - R(\delta)\right\}. \tag{4.36}$$

We now turn to the parabolic problem when $J < \infty$. Estimates are more cumbersome because the dead core may not be monotonic in t. For instance, if $C_0(x) = 1$, a dead core may form after a certain time and later disappear. We can, however, find an upper bound on T, the time beyond which $C(t,x) > 0$ for all $x \in \bar{\Omega}$. We confine ourselves to the Dirichlet problem ($\alpha = 0$).

THEOREM 4.6. *Let $\alpha = 0, J < \infty$, and let a be the half-width of the thinnest strip enclosing Ω. Then*

$$T \le \frac{\varepsilon + \lambda}{2} a^2, \qquad \hat{T} \le \frac{\lambda}{2} a^2. \tag{4.37}$$

Our proof is based on comparison with a half-space, so we begin with the following lemma.

LEMMA 4.1. *Consider problem* (P′) *for the half-space $x > 0$ with $C(t,0) = 1$ for $t > 0$ and initial values $0 \le X(0,x) \le 1, 0 \le C(0,x) \le 1$. Let*

$$\rho(t) = \inf\{x : C(t,x) \equiv 0\}$$

be the penetration distance of the gas, where $\rho(t)$ is understood to be $+\infty$ if $C(x,t) > 0$ for all x. Then

$$\rho^2 \geq \frac{2t}{\varepsilon + \lambda}, \tag{4.38}$$

an estimate which holds even when $\varepsilon = 0$.

Proof of Lemma 4.1. We give the proof for $\varepsilon > 0$, omitting the simpler case $\varepsilon = 0$. Let us first consider the problem with *zero initial values* for C and X. Condition (3.2) being satisfied, $C(t,\cdot)$ and $\rho(t)$ are monotonically increasing. We write (2.3) as

$$\varepsilon C_t - C_{xx} = -\lambda X_t, \quad x > 0, \quad t > 0$$

and, as in [14], integrate in time from 0 to t to obtain

$$\varepsilon C(t,x) + \lambda X(t,x) = \psi_{xx}, \tag{4.39}$$

where

$$\psi(t,x) = \int_0^t C(\tau,x)\,d\tau.$$

We multiply (4.39) by x and integrate from $x = 0$ to $x = \rho(t)$:

$$\int_0^\rho x(\varepsilon C + \lambda X)\,dx = \int_0^\rho x\psi_{xx}\,dx = t - \psi(t,\rho(t)).$$

By the time monotonicity, $C(\tau,\rho(t)) = 0$ for $\tau < t$, so that $\psi(t,\rho(t)) = 0$. Thus we find

$$t = \int_0^\rho x[\varepsilon C + \lambda X]\,dx \leq (\varepsilon + \lambda)\frac{\rho^2}{2},$$

which proves (4.38) for zero initial values. For other initial values the gas concentration $C(t,x)$ will be larger (Property I, §3.1), and (4.38) must remain true. □

Proof of Theorem 4.6. Let (X_Ω, C_Ω) be the solution of the problem (P′) with $\alpha = 0$. We compare this solution with the solution (X_H, C_H) for a supporting half-space H enclosing Ω with initial values $(X_H(0,x), C_H(0,x)) \leq (X_\Omega(0,x), C_\Omega(0,x))$. Then, since $C_H \leq 1$ on $\partial\Omega$, $(X_H(t,x), C_H(t,x))$ is a subsolution to (P′) for Ω, and hence

$$C_\Omega(t,x) \geq C_H(t,x). \tag{4.40}$$

Now let U be the thinnest slab enclosing Ω; applying (4.40) successively to the half-spaces corresponding to the two faces of U, we see that $C_\Omega(t,x) > 0$ when $t > \frac{\varepsilon+\lambda}{2}a^2$, where a is the half-width of the slab. This yields (4.37). The proof is the same for the pseudo-steady-state case. □

Remark 4.8. When full conversion occurs in finite time ($I < \infty$), we have estimated t_1 and $\hat{t}_1$, the times to full conversion. When $t > t_1$ (or $t > \hat{t}_1$ in the p.s.s. case), the equation for the gas concentration is just the ordinary heat equation, which has the well-known property $C > 0$ for the given boundary conditions. Hence t_1 is an upper bound for T.

We end this section by showing that a dead core can occur for the half-space problem when $g(C) = C^p, p < 1$. Consider again the Dirichlet problem, now with

$C(0,x) = S(0,x) = 1$, the initial condition on C being least favorable for generating a dead core. Our problem then becomes

$$\begin{aligned} &\varepsilon C_t - C_{xx} = -\lambda f(S)g(C) = \lambda S_t, \quad x > 0, \quad t > 0; \\ &C(0,x) = S(0,x) = 1, \qquad C(t,0) = 1. \end{aligned} \tag{4.41}$$

Since conversion of the solid is fastest at $x = 0$, we have

$$f(S(t,0)) \leq f(S(t,x)) \leq 1.$$

From the second inequality we find that $C(t,x) \geq z(t)$, where $z(t)$ is the solution of $z_t = -\lambda g(z), z(0) = 1$. Because $z > 0$ for

$$t < K \doteq \int_0^1 \frac{dz}{g(z)},$$

we see that, as expected, $C(t,x)$ is strictly positive for $t < K$. Note that when $g(z) = z^p, K$ is finite if and only if $p < 1$. Next we show that C develops a dead core at later times by constructing a supersolution $D(t,x)$ of (4.41) over a bounded time interval with $D(t,x) = 0$ for x sufficiently large and some range of t. Since $S(t,0)$ satisfies $S_t = -f(S)$ with initial value one, we find that $S(t,0) = R^{-1}(t)$ (see (4.5) and (4.6)). Because $S(t,0)$ decreases monotonically from 1 to 0, we can find, for each δ with $0 < \delta < 1$, a time T_δ such that

$$f(S(t,0)) \geq 1 - \delta, \qquad 0 < t < T_\delta.$$

Therefore, $f(S(t,x)) \geq 1 - \delta$ for $0 < t < T_\delta$ and $C(t,x) \leq D(t,x)$ on $(0,T_\delta)$ where D satisfies the scalar problem

$$\begin{cases} \varepsilon D_t - D_{xx} = -\lambda(1-\delta)g(D), \quad x > 0, \quad t > 0; \\ D(0,x) = 1, \qquad D(t,0) = 1. \end{cases} \tag{4.42}$$

It was shown in [32] that problem (4.42) exhibits a dead core for all $t > \frac{\varepsilon K}{\lambda(1-\delta)}$. Therefore, (4.41) will have a dead core if we can choose the parameters so that

$$\frac{\varepsilon K}{\lambda(1-\delta)} < T_\delta. \tag{4.43}$$

Let us illustrate the calculation for the case $f(S) = S$. Then $S(t,0) = e^{-t}, T_\delta = -\log(1-\delta)$, and (4.43) is satisfied if, for some $\delta, 0 < \delta < 1$, we have

$$\frac{\varepsilon K}{\lambda} < -(1-\delta)\log(1-\delta).$$

The maximum of the right side occurs at $\delta = 1 - \frac{1}{e}$, and the maximal value is 1. Thus (4.41) will have a dead core if $\frac{\varepsilon K}{\lambda} < 1$.

Remark 4.9. A *bounded* domain Ω of sufficiently large size will also have a dead core for a suitable choice of the parameters. Of course, in this case the dead core must disappear in finite time.

REFERENCES

[1] R. ARIS, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts,* Clarendon Press, Oxford, 1975.

[2] C. BANDLE, *Isoperimetric Inequalities and Their Applications,* Pitman, London, 1980.

[3] C. BANDLE, R. SPERB, AND I. STAKGOLD, *Diffusion-reaction with monotone kinetics,* Nonlinear Anal., 8 (1984), pp. 321–333.

[4] C. BANDLE AND I. STAKGOLD, *The formation of the dead core in parabolic reaction-diffusion equations,* Trans. Amer. Math. Soc., 286 (1984), pp. 275–293.

[5] M. BERTSCH, R. KERSNER, AND L. A. PELETIER, *Positivity versus localization in degenerate diffusion problems,* Nonlinear Anal., 9 (1985), pp. 831–847.

[6] D. BLANCHARD, A. DAMLAMIAN, AND H. GHIDOUCHE, *A nonlinear system for phase change with dissipation,* Differential and Integral Equations, 2 (1989), pp. 344–362.

[7] J. I. DIAZ, *Nonlinear Partial Differential Equations and Free Boundaries, Vol. 1. Elliptic Equations,* Pitman, London, 1985.

[8] J. I. DIAZ AND J. HERNÁNDEZ, *On the existence of a free boundary for a class of reaction diffusion systems,* SIAM J. Math. Anal., 15 (1984), pp 670–685.

[9] J. I. DIAZ AND J. HERNÁNDEZ, *Some results on the existence of free boundaries for parabolic reaction-diffusion systems,* In Trends in Theory and Practice of Nonlinear Differential Equations, V. Lakshmikantham, ed. Marcel Dekker, 1984, pp. 149–156.

[10] J. I. DIAZ AND I. STAKGOLD, *Mathematical analysis of the conversion of a porous solid by a distributed gas reaction,* in Proc. of the XI CEDYA, Univ. of Malaga, Spain 1989, pp. 217–223.

[11] J. I. DIAZ AND I. I. VRABIE, *Existence for reaction diffusion systems, A compactness method approach,* submitted.

[12] A. DI LIDDO AND L. MADDALENA, *Mathematical Analysis of a Chemical Reaction with a Lumped Temperature and Strong Absorption,* J. Math. Anal. Appl., 163 (1992), pp. 86–102.

[13] A. DI LIDDO, L. MADDALENA, AND I. STAKGOLD, *Traveling waves for distributed gas-solid reactions,* J. Differential Equations, to appear.

[14] A. DI LIDDO AND I. STAKGOLD, *Isothermal combustion with two moving fronts,* J. Math. Anal. Appl., 152 (1990), pp. 584–599.

[15] A. FRIEDMAN AND A. E. TZAVARAS, *A quasilinear parabolic system arising in modelling of catalytic reactors,* J. Differential Equations, 70 (1987), pp. 167–196.

[16] G. F. FROMENT AND K. B. BISCHOFF, *Chemical Reactor Analysis and Design,* Wiley, New York 1979.

[17] J. GRAHAM-EAGLE AND I. STAKGOLD, *A steady-state diffusion problem with fractional power absorption rate,* IMA J. Appl. Math., 39 (1987), pp. 67–73.

[18] G. S. LADDE, V. LAKSHMIKANTHAM, AND A. S. VATSALA, *Monotone Iterative Techniques for Nonlinear Differential Equations,* Pitman, London, 1985.

[19] A. MCNABB, *Asymptotic behaviour of solutions of diffusion equations,* J. Math. Anal. Appl., 51 (1975), pp. 219–222.

[20] A. MCNABB AND G. KEADY, *Some explicit solutions of $-\Delta w = 1$ with zero boundary data,* Report 42, Dept. of Mathematics, The University of Western Australia, Nedlands, Australia, 1988.

[21] C. V. PAO, *Asymptotic behavior and nonexistence of global solutions for a class of nonlinear boundary value problems of parabolic type,* J. Math. Anal. Appl., 65 (1978), pp. 616–637.

[22] ———, *On nonlinear reaction-diffusion systems,* J. Math. Anal. Appl., 87 (1982), pp. 165–198.

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

[24] I. STAKGOLD, *Gas-solid reactions,* in Dynamical Systems II, A. R. Bednarek and L. Cesari, eds., Academic Press, New York, 1982, pp. 403–417.

[25] ———, *Partial extinction in reaction-diffusion,* Conferenze del Seminario di Matematica, Universitá di Bari, Bari, Italy, 224, 1987.

[26] ———, *Localization and extinction in reaction-diffusion,* in Free Boundary Problems: Theory and Applications, Vol. 1, K. H. Hoffmann et al., eds., Longman, London, 1988, pp. 208–221.

[27] I. STAKGOLD, K. B. BISCHOFF, AND V. GOKHALE, *Validity of the pseudo-steady-state approximation,* Internat. J. Eng. Sci., 21 (1983), pp. 537–542.

[28] I. STAKGOLD AND A. MCNABB, *Conversion estimates for gas-solid reactions,* Math. Modelling, 5 (1984), pp. 325–330.
[29] J. SZEKELY, J. W. EVANS, AND H. Y. SOHN, *Gas-solid reactions,* Academic Press, New York, 1976.
[30] J. L. VÁZQUEZ, *A strong maximum principle for some quasilinear elliptic equations,* Appl. Math. Optim., 12 (1984), pp. 191–202.
[31] I. I. VRABIE, *Compactness Methods for Nonlinear Evolutions,* Longman, London, 1987.
[32] C. BANDLE AND I. STAKGOLD, *Reaction-Diffusion and Dead Cores,* in Free Boundary Problems: Applications and Theory, Vol. IV, Pitman, London, 1985.
[33] R. MARTIN, *Mathematical models in gas-liquid reactions,* Nonlinear Anal., Theory and Appl., 4 (1980), pp. 509–527.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 329–339, March 1995

004

POSITIVE SOLUTIONS OF SINGULAR SUBLINEAR DIRICHLET BOUNDARY VALUE PROBLEMS *

YONG ZHANG†

Abstract. This paper mainly studies the existence of positive solutions of Dirichlet boundary value problems for a class of singular sublinear ordinary differential equations. A necessary and sufficient condition for the existence of $C[0,1]$ positive solutions as well as $C^1[0,1]$ positive solutions is given. The uniqueness of the solution is also concerned, and an application to the Dirichlet problem of semilinear elliptic equations is given.

Key words. sublinear equations, singular Dirichlet problems, positive solutions, lower and upper solutions, radial solutions of semilinear elliptic equations, existence and uniqueness

AMS subject classifications. primary 34B15; secondary 35J65

1. Introduction.

In this paper we are concerned about the singular boundary value problem of second order ordinary differential equations

$$\ddot{x} + f(t,x) = 0, \qquad t \in (0,1), \tag{1.1}$$

$$x(0) = x(1) = 0. \tag{1.2}$$

By singularity we mean that the function f in (1.1) is allowed to be unbounded at the end points $t = 0$ and 1. Recently such problems have interested many authors [1]–[8]. For background, we mention the existence problem of radial solutions of nonlinear elliptic equations as follows:

$$\Delta u + f(|x|,u) = 0, \qquad x \in \Omega,$$
$$u(x)|_{\partial\Omega} = 0, \qquad |x|^{n-1}\left.\frac{\partial u}{\partial \vec{n}}\right|_{\partial\Omega} \text{ exists},$$

where Ω is an n-ball of radius ρ centered at $0 \in R^n, 0 < \rho \le +\infty, n \ge 2$, and when $\rho = +\infty$ (i.e., $\Omega = R^n$) we assume $n > 2$; $f(r,u)$ is continuous for $0 \le r < \rho, u \in (-\infty,\infty)$ (or $(0,\infty)$ if we consider positive solutions).

This problem leads one to treat the boundary value problem of ordinary differential equations

$$\frac{1}{p(r)}\frac{d}{dr}\left(p(r)\frac{du}{dr}\right) + f(r,u) = 0, \qquad r \in (0,\rho),$$
$$\dot{u}(0) = 0, \quad u(\rho) = 0 \quad \text{and} \quad p(\rho)\dot{u}(\rho) \text{ exists},$$

where $p(r) = r^{n-1}$. Under the transformation $t = t(r) = (\int_r^\rho (ds/p(s)) + 1)^{-1}, y(t) = tu(r)$, the above problem is reduced to a singular Dirichlet problem of the form (1.1), (1.2), that is,

$$\ddot{y}(t) + (p^2(r(t))/t^3) f(r(t), y/t) = 0, \qquad t \in (0,1),$$
$$y(0) = y(1) = 0, \quad y(t) \in C^1[0,1] \cap C^2(0,1),$$

* Received by the editors March 23, 1993; accepted for publication (in revised form) September 28, 1993.

† Department of Mathematics and Astronomy, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada.

where $r(t)$ is the inverse of $t(r)$.

To treat the singular problem (1.1), (1.2), we can make use of three general methods, the shooting method, the operator or topological degree method, and the lower and upper solution method. The first has been used successfully in the study of some special singular problems such as negative exponent Emden–Fowler boundary problems (cf. [1]). In fact, this method is available if one assumes that $f(t,x)$ is decreasing in x. But for other cases it often seems useless. The second method, though it has many advantages in treating nonsingular problems, still has some difficulties when treating singular problems. One can find some results concerning this method for singular problems in [2], [3], and their references. Other valuable results connected with the operator iteration technique are obtained in [4] and [5], where singular boundary value problems with more general boundary conditions have been treated under the condition that $f(t,x)$ is decreasing in x (see [6] and [7] for relevant results). The third method was developed originally for nonsingular problems. Since there are many differences between singular and nonsingular problems, using it as a general method to the singular problems is pending more basic investigations. Some attempts have been made by Zhang in [8] to treat the singular problem (1.1), (1.2) with $f(t,x) = p(t)x^{\alpha}$, where $p(t) \in C(0,1)$, $p(t) > 0, 0 < \alpha < 1$, and in [9] to treat singular boundary value problems of other forms. These and the discussion in this paper show that the lower and upper solution technique is really a very promising method for the study of singular boundary value problems.

In this paper we deal mainly with the existence of positive solutions of the singular boundary value problem (1.1), (1.2) under the following sublinear hypothesis.

(E) $f(t,x) \in C((0,1)\times(0,\infty),[0,\infty))$, $f(t,1) \not\equiv 0$ for $t \in (0,1)$, and there exist constants $\lambda, \mu, N, M (-\infty < \lambda \le \mu < 1, 0 < N \le 1 \le M)$, such that, for $t \in (0,1)$ and $x \in (0,\infty)$,

$$c^{\mu} f(t,x) \le f(t,cx) \le c^{\lambda} f(t,x), \quad \text{if } 0 \le c < N; \tag{1.3}$$

$$c^{\lambda} f(t,x) \le f(t,cx) \le c^{\mu} f(t,x), \quad \text{if } c \ge M. \tag{1.4}$$

Typical functions that satisfy the above sublinear hypothesis are those taking the form

$$f(t,x) = \sum_{k=1}^{n} p_k(t) x^{\lambda_k};$$

here $p_k(t) \in C(0,1)$, $p_k(t) > 0$ on $(0,1)$, $\lambda_k < 1, k = 1,2,\dots,n$.

We will call a function $x(t) \in C[0,1] \cap C^2(0,1)$ a $C[0,1]$ (positive) solution of (1.1), (1.2) if it satisfies (1.1) on $(0,1)$ and the boundary condition (1.2) (and $x(t) > 0$ for $t \in (0,1)$). If, in addition, $\dot{x}(0+)$ and $\dot{x}(1-)$ exist, i.e., $x(t) \in C^1[0,1] \cap C^2(0,1)$, we will call it a $C^1[0,1]$ (positive) solution. Here we point out that, under the condition (E), any nontrivial nonnegative solution of (1.1), (1.2) is in fact a positive solution, since it is a concave function.

The main results of this paper are the following two theorems which will be proved in §3.

THEOREM 1. *Suppose* (E) *holds. Then a necessary and sufficient condition for problem* (1.1), (1.2) *to have a* $C[0,1]$ *positive solution is that*

$$\int_0^1 t(1-t) f(t,1)\, dt < \infty. \tag{1.5}$$

THEOREM 2. *Suppose* (E) *holds. Then a necessary and sufficient condition for problem* (1.1), (1.2) *to have a* $C^1[0,1]$ *positive solution is that*

$$\int_0^1 f(t, t(1-t))\,dt < \infty. \tag{1.6}$$

With respect to relevant results, the author would like to point out that a condition similar to (1.6) has been used by Gatica, Oliker, and Waltman in [4] to obtain $C^1[0,1]$ solutions to singular problems with mixed boundary conditions. The assumption that $f(t,x)$ is decreasing in x is a crucial condition there. Within the scope of Dirichlet problems, the sufficiency part of our Theorem 2 extends their Theorem 2.2, but does not include it.

Though the uniqueness seems to be a property typical of sublinear problems, fully studying it is not easy, especially when $f(t,x)$ has no monotone property in x. We will discuss the uniqueness in §4, where we will mainly prove the following uniqueness theorem.

THEOREM 3. *Suppose that*

(U) $f(t,x) \in C((0,1)\times(0,\infty),[0,\infty))$, *and* $f(t,x)/x$ *is strictly decreasing in* x *for all* $t \in (0,1)$.

Then, if problem (1.1), (1.2) *has a* $C^1[0,1]$ *positive solution, it will admit no other positive solutions.*

As an application of the above theorems we will discuss the existence and uniqueness of radial solutions to Dirichlet problems of semilinear elliptic equations in §5.

2. Preliminary lemmas.

LEMMA 1. *For any* $f(t) \in L^1[0,1]$, *there exists* $g(t) \in C[0,1]$ *with* $g(t) > 0$ *for* $t \in (0,1)$ *and* $g(0) = g(1) = 0$, *such that* $f(t)/g(t) \in L^1[0,1]$.

Proof. Suppose $f(t) \in L^1[0,1]$. Without loss of generality we can assume $f(t) > 0$ for $t \in [0,1]$. Let $l = \int_0^1 f(t)\,dt$, $h(t) = \int_0^t f(s)\,ds$, and, for any given $\varepsilon : 0 < \varepsilon < 1$, let

$$g(t) = [h(t)(l - h(t))]^{\varepsilon}, \qquad t \in [0,1].$$

Then $g(t) \in C[0,1]$, $g(t) > 0$ for $t \in (0,1)$, and $g(0) = g(1) = 0$. Also,

$$0 < \int_0^1 (f(t)/g(t))\,dt = \int_0^l [h(l-h)]^{-\varepsilon}\,dh < \infty.$$

This shows $f(t)/g(t) \in L^1[0,1]$ and hence completes the proof. □

LEMMA 2. *For any* $g(t) \in C[0,1]$ *such that* $g(0) = g(1) = 0$ *and* $g(t) \geq 0$ *for* $t \in [0,1]$, *there exists* $q(t) \in C[0,1] \cap C^2(0,1)$ *such that* $q(0) = q(1) = 0, q(t) \geq g(t)$, *and* $q''(t) \leq 0$ *for* $t \in (0,1)$.

The proof of Lemma 2 will be given in the appendix because of the lengthy argument there involved.

Now let us consider the general singular boundary value problems as follows:

$$\ddot{x} + f(t,x) = 0, \qquad t \in (0,1), \tag{2.1}$$

$$x(0) = r_1, \qquad x(1) = r_2, \tag{2.2}$$

where we assume $f(t,x) \in C((0,1)\times I, R), r_1, r_2 \in I, I$ is an interval in R.

A function $\alpha(t)$ is called a lower solution of (2.1) if $\alpha(t) \in C^2(0,1)$ and, for $t \in (0,1), \alpha(t) \in I, \ddot{\alpha}(t) + f(t,\alpha(t)) \geq 0$. Similarly, $\beta(t)$ is called an upper solution of (2.1) if $\beta(t) \in C^2(0,1)$ and, for $t \in (0,1), \beta(t) \in I, \ddot{\beta}(t) + f(t,\beta(t)) \leq 0$.

LEMMA 3. *Assume that there exist lower and upper solutions of* (2.1), *respectively* $\alpha(t)$ *and* $\beta(t)$, *such that* $\alpha(t), \beta(t) \in C[0,1] \cap C^2(0,1), \alpha(t) \leq \beta(t)$ *for* $t \in [0,1], \alpha(0) = \beta(0) = r_1$, *and* $\alpha(1) = \beta(1) = r_2$. *Then problem* (2.1), (2.2) *admits at least one* $C[0,1]$ *solution* $x(t)$ *such that* $\alpha(t) \leq x(t) \leq \beta(t)$ *for* $t \in [0,1]$. *If, in addition, there exists* $F(t) \in L^1[0,1]$ *such that*

$$|f(t,x)| \leq F(t), \qquad t \in (0,1), \quad \alpha(t) \leq x \leq \beta(t), \tag{2.3}$$

then the solution $x(t)$ *is a* $C^1[0,1]$ *solution.*

Proof. Let $\{a_n\}, \{b_n\}$ be sequences satisfying $0 < \cdots < a_{n+1} < a_n < \cdots < a_1 < b_1 < \cdots < b_n < b_{n+1} < \cdots < 1, a_n \to 0$ and $b_n \to 1$ as $n \to \infty$, and let $\{r_1^{(n)}\}, \{r_2^{(n)}\}$ be sequences satisfying $\alpha(a_n) \leq r_1^{(n)} \leq \beta(a_n), \alpha(b_n) \leq r_2^{(n)} \leq \beta(b_n), n = 1, 2, \ldots$. For each n consider the nonsingular problem

$$\ddot{x} + f(t,x) = 0, \tag{2.1}$$

$$x(a_n) = r_1^{(n)}, \qquad x(b_n) = r_2^{(n)}. \tag{2.4}$$

By a well-known corresponding result in nonsingular boundary value problems (see [13, Thm. 1.5.1]), we come to the conclusion that, for each n, the problem (2.1), (2.4) admits a solution $x_n(t) \in C^2[a_n, b_n]$ such that $\alpha(t) \leq x_n(t) \leq \beta(t), t \in [a_n, b_n]$. Since $[a_1, b_1] \subset [a_n, b_n], n = 1, 2, \ldots$, there is, for each $n, t_n \in [a_1, b_1]$ such that $|\dot{x}_n(t_n)| = |(x_n(b_1) - x_n(a_1))/(b_1 - a_1)| \leq (2/(b_1 - a_1)) \max_{t \in [a_1, b_1]}(|\alpha(t)| + |\beta(t)|)$. This allows us to assume (substituting by subsequences if necessary)

$$t_n \to t_0 \in [a_1, b_1], \quad x_n(t_n) \to x_0 \in [\alpha(t_0), \beta(t_0)], \quad x_n(t_n) \to x_0' \in R, \quad \text{as } n \to \infty.$$

From [14, Thm. 3.2, p. 14] there is a solution $x(t)$ of (2.1), with the maximum existence interval (ω^-, ω^+), such that $x(t_0) = x_0, \dot{x}(t_0) = x_0'$, and $x_n(t)$ converges uniformly to $x(t)$ on any compact subinterval of (ω^-, ω^+). Since

$$[a_n, b_n] \subset [a_{n+1}, b_{n+1}], \quad \bigcup_{n=1}^{\infty} [a_n, b_n] = (0,1) \quad \text{and} \quad \alpha(t) \leq x_n(t) \leq \beta(t) \quad \text{for} \quad t \in [a_n, b_n],$$

one can easily see that $\alpha(t) \leq x(t) \leq \beta(t)$ for $t \in (\omega^-, \omega^+)$. This leads additionally to the fact that $(\omega^-, \omega^+) = (0,1)$, from the Extension Theorem. Also, $x(t)$ satisfies (2.2) because $\alpha(t)$ and $\beta(t)$ do. Thus, $x(t)$ is a $C[0,1]$ solution of (2.1), (2.2).

In addition, if (2.3) holds, then $|\ddot{x}(t)| \leq F(t)$, and hence $\ddot{x}(t)$ is absolutely integrable on $[0,1]$. This implies $x(t) \in C^1[0,1]$, so $x(t)$ is a $C^1[0,1]$ solution of the problem (2.1), (2.2). The proof is complete. □

3. Proof of the main existence results. It is easy to check that in the hypothesis (E) the number λ, μ can be assumed, without loss of generality, to satisfy $\lambda < 0 < \mu < 1$. This will be supposed throughout this section for convenience.

Proof of Theorem 1. 1°. Necessity. Assume that $x(t)$ is a $C[0,1]$ positive solution of problem (1.1), (1.2). Then there is a t_0 such that $\dot{x}(t_0) = 0$, and hence

$$-\dot{x}(t) = \int_{t_0}^{t} f(s, x(s))\, ds, \qquad t \in (0,1). \tag{3.1}$$

Let $c > 0$ be a constant such that $cx(t) \le N$ for $t \in [0,1]$ and $1/c \ge M$. From (1.3) and (1.4) we have

$$f(t,x(t)) \ge (1/c)^{\lambda} f(t,cx(t)) \ge c^{\mu-\lambda} x^{\mu}(t) f(t,1), \qquad t \in (0,1). \tag{3.2}$$

(3.1) as well as condition (E) implies that $\dot{x}(t) \ge 0$ (i.e., $x(t)$ increases) for $t \in (0,t_0)$, and $\dot{x}(t) \le 0$ (i.e., $x(t)$ decreases) for $t \in (t_0,1)$. This, combined with (3.2), yields

$$\begin{aligned} \dot{x}(t) = \int_t^{t_0} f(s,x(s))\,ds &\ge c^{\mu-\lambda} \int_t^{t_0} x^{\mu}(s) f(s,1)\,ds \\ &\ge c^{\mu-\lambda} x^{\mu}(t) \int_t^{t_0} f(s,1)\,ds, \qquad t \in (0,t_0), \end{aligned}$$

and

$$-\dot{x}(t) = \int_{t_0}^{t} f(s,x(s))\,ds \ge c^{\mu-\lambda} x^{\mu}(t) \int_{t_0}^{t} f(s,1)\,ds, \qquad t \in (t_0,1).$$

Dividing both sides of each of the above two inequalities by $c^{\mu-\lambda}x^{\mu}(t)$ and then integrating them, respectively, on $[0,t_0]$ and $[t_0,1]$ we have

$$\begin{aligned} 0 < \int_0^{t_0} t f(t,1)\,dt &= \int_0^{t_0} dt \int_t^{t_0} f(s,1)\,ds \le c^{\lambda-\mu} x^{1-\mu}(t_0)/(1-\mu) < \infty, \\ 0 < \int_0^{1} (1-t) f(t,1)\,dt &= \int_{t_0}^{1} dt \int_{t_0}^{t} f(s,1)\,ds \le c^{\lambda-\mu} x^{1-\mu}(t_0)/(1-\mu) < \infty. \end{aligned}$$

These imply that (1.5) holds.

2°. Sufficiency. Suppose (1.5) holds. Then, by Lemmas 1 and 2, there exists $q(t) \in C[0,1] \cap C^2(0,1)$ satisfying $q(t) > 0$ and $q''(t) \le 0$ for $t \in (0,1)$, and $q(0) = q(1) = 0$, such that

$$\int_0^1 t(1-t) q^{\lambda-\mu}(t) f(t,1)\,dt < \infty. \tag{3.3}$$

This, together with (1.3) and (1.4), implies

$$\int_0^1 t(1-t) q^{-\mu}(t) f(t,q(t))\,dt < \infty. \tag{3.4}$$

Let

$$\begin{aligned} \Gamma_1 &= (1-t) \int_0^t s^{1+\mu}(1-s)^{\mu} f(s,1)\,ds + t \int_t^1 s^{\mu}(1-s)^{1+\mu} f(s,1)\,ds, \\ \Gamma_2 &= (1-t) \int_0^t s q^{-\mu}(s) f(s,q(s))\,ds + t \int_t^1 (1-s) q^{-\mu}(s) f(s,q(s))\,ds + q(t). \end{aligned}$$

One can check that $\Gamma_i \in C[0,1] \cap C^2(0,1), \Gamma_i(0) = \Gamma_i(1) = 0, i = 1,2$, and

$$L_1 t(1-t) \le \Gamma_1(t) \le L_1, \quad q(t) \le \Gamma_2(t) \le L_2, \quad t \in [0,1];$$

$$\Gamma_1''(t) = -t^{\mu}(1-t)^{\mu} f(t,1), \quad \Gamma_2''(t) \le -q^{-\mu}(t) f(t,q(t)), \quad t \in (0,1).$$

Here

$$L_1 = \int_0^1 s^{1+\mu}(1-s)^{1+\mu} f(s,1)\,ds,$$

$$L_2 = \int_0^1 s(1-s)q^{-\mu}(s) f(s,q(s))\,ds + q_0, \qquad q_0 = \max_{t\in[0,1]} q(t).$$

Let $\alpha(t) = a\Gamma_1(t), \beta(t) = b\Gamma_2(t), t \in [0,1]$; here a, b are constants satisfying $0 < a \le 1 \le b$, and their sizes will be determined later. Suppose c_1, c_2 are constants such that $c_1 L_1 \le N, 1/c_1 \ge M, c_2 \ge M, 1/c_2 \le N$. From (1.3), (1.4) we then have

$$\begin{aligned} f(t,\alpha(t)) &\ge (1/c_1)^\lambda f(t, c_1\alpha(t)) \ge c_1^{\mu-\lambda}\alpha^\mu(t) f(t,1) \\ &\ge a^\mu L_1^\mu c_1^{\mu-\lambda} t^\mu (1-t)^\mu f(t,1), \qquad t \in (0,1), \\ f(t,\beta(t)) &\le c_2^{\mu-\lambda}(\beta(t)/q(t))^\mu f(t,q(t)) \\ &\le b^\mu L_2^\mu c_2^{\mu-\lambda} q^{-\mu}(t) f(t,q(t)), \qquad t \in (0,1). \end{aligned}$$

Again, according to (1.3), (1.4) we can find a $k_0 > 0$ such that $f(t,q(t)) \ge k_0 q^\mu(t)$ $f(t,1)$, and consequently, from the definitions of $\Gamma_1(t), \Gamma_2(t)$, we have, when $k > k_0^{-1}, \Gamma_1(t) \le k\Gamma_2(t)$ for $t \in [0,1]$.

Now we choose $a = \min\{1, (L_1^\mu c_1^{\mu-\lambda})^{1/(1-\mu)}\}$ and $b = \max\{1, k_0^{-1}, (L_2^\mu c_2^{\mu-\lambda})^{1/(1-\mu)}\}$. Then the above discussions show that, for such choice of a and $b, \alpha(t)$ and $\beta(t)$ are lower and upper solutions of (1.1), respectively, and satisfy $0 < \alpha(t) \le \beta(t)$ for $t \in (0,1)$ and $\alpha(i) = \beta(i) = 0$ for $i = 0, 1$. From the first conclusion of Lemma 3 we deduce that the problem (1.1), (1.2) admits a $C[0,1]$ solution $x(t)$ satisfying $0 < \alpha(t) \le x(t) \le \beta(t)$ for $t \in (0,1)$. This completes the proof. □

Proof of Theorem 2. 1°. Necessity. Assume that $x(t)$ is a $C^1[0,1]$ positive solution of (1.1), (1.2). Then $\dot{x}(0) > 0$ and $\dot{x}(1) < 0$ since (1.2) holds and $\ddot{x}(t) \le 0, x(t) > 0$ for $t \in (0,1)$. This implies that there are constants I_1 and $I_2, 0 < I_1 < I_2$, such that

$$I_1 t(1-t) \le x(t) \le I_2 t(1-t), \qquad t \in [0,1]. \tag{3.5}$$

Let c be a constant satisfying $cI_2 \le N, 1/c \ge M$. Then (1.3), (1.4), and (3.5) lead to

$$\begin{aligned} f(t,x(t)) &\ge (1/c)^\lambda f\left(t, \frac{cx(t)}{t(1-t)} t(1-t)\right) \\ &\ge c^{\mu-\lambda}(x(t)/t(1-t))^\mu f(t, t(1-t)) \\ &\ge c^{\mu-\lambda} I_1^\mu f(t, t(1-t)), \qquad t \in (0,1). \end{aligned}$$

Dividing by $c^{\mu-\lambda} I_1^\mu$ and integrating, we get

$$\begin{aligned} \int_0^1 f(t, t(1-t))\,dt &\le c^{\lambda-\mu} I_1^{-\mu} \int_0^1 f(t,x(t))\,dt \\ &= c^{\lambda-\mu} I_1^{-\mu}(\dot{x}(0) - \dot{x}(1)) < \infty. \end{aligned}$$

Thus (1.6) has been derived.

2°. Sufficiency. Suppose that (1.6) holds. Let

$$\Gamma(t) = (1-t)\int_0^t s f(s, s(1-s))\,ds + t\int_t^1 (1-s) f(s, s(1-s))\,ds.$$

Then $\Gamma(t) \in C[0,1] \cap C^2(0,1)$ and (3.5) holds if $x(t)$ is replaced by $\Gamma(t)$ and $I_1 = \int_0^1 t(1-t)f(t,t(1-t))\,dt, I_2 = \int_0^1 f(t,t(1-t))\,dt$. Let $\alpha(t) = a\Gamma(t), \beta(t) = b\Gamma(t)$; here $a = \min\{1, (\bar{c}^{\lambda-\mu}I_2^{\lambda})^{1/(1-\mu)}\}, b = \max\{1, (\bar{c}^{\mu-\lambda}I_2^{\mu})^{1/(1-\mu)}\}, \bar{c}$ is a constant satisfying $\bar{c}I_1 \geq M, 1/\bar{c} \leq N$. A similar argument to that we have clarified in part 2° of the proof of Theorem 1 yields that, for $t \in (0,1)$,

$$\begin{aligned}
\ddot{a}(t) + f(t,\alpha(t)) &= -af(t,t(1-t)) + f(t,a\Gamma(t)) \\
&\geq -af(t,t(1-t)) + a^{\mu}\bar{c}^{\lambda-\mu}I_2^{\lambda}f(t,t(1-t)) \geq 0, \\
\ddot{\beta}(t) + f(t,\beta(t)) &= -bf(t,t(1-t)) + f(t,b\Gamma(t)) \\
&\leq -bf(t,t(1-t)) + b^{\mu}\bar{c}^{\mu-\lambda}I_2^{\mu}f(t,t(1-t)) \leq 0.
\end{aligned}$$

So $\alpha(t), \beta(t)$ are, respectively, lower and upper solutions of (1.1), satisfying $0 < \alpha(t) \leq \beta(t)$ for $t \in (0,1)$ and $\alpha(i) = \beta(i) = 0, i = 0,1$. Additionally, when $t \in (0,1)$ and $\alpha(t) \leq x \leq \beta(t)$,

$$\begin{aligned}
0 \leq f(t,x) &\leq (a/\bar{c})^{\lambda} f\left(t, \frac{\bar{c}x}{at(1-t)}t(1-t)\right) \\
&\leq (a/\bar{c})^{\lambda}\left(\frac{\bar{c}x}{at(1-t)}\right)^{\mu} f(t,t(1-t)) \\
&\leq (a/\bar{c})^{\lambda-\mu}(bI_2)^{\mu}f(t,t(1-t)) =: F(t).
\end{aligned}$$

From (1.6) we assert $\int_0^1 F(t)\,dt < \infty$. Thus, according to Lemma 3, we have proved that the problem (1.1), (1.2) admits a $C^1[0,1]$ solution $x(t)$ such that $\alpha(t) \leq x(t) \leq \beta(t)$ for $t \in [0,1]$. This completes the proof. □

4. Uniqueness. In this section we make a brief discussion on the uniqueness of the solution of problem (1.1), (1.2). We can assert the uniqueness for two special cases. The first is that $f(t,x)$ is decreasing in x; the second is that $f(t,x)/x$ is decreasing in x. In the former case the uniqueness is easy to obtain through a standard argument. In fact, in this case, for two solutions of (1.1), say $x_1(t), x_2(t)$, with $x_1(t_i) = x_2(t_i), i = 1,2, x_1(t) \leq x_2(t)$ for $t \in (t_1,t_2)$, where $0 \leq t_1 \leq t_2 \leq 1$, it must be true that $\lim_{t\to t_1}\sup(x_2'(t) - x_1'(t)) \geq 0, \lim_{t\to t_2}\inf(x_2'(t) - x_1'(t)) \leq 0$ and $x_2''(t) - x_1''(t) = f(t,x_1(t)) - f(t,x_2(t)) \geq 0$ for $t \in (t_1,t_2)$, which assure one that $x_1(t) \equiv x_2(t)$ for $t \in [t_1,t_2]$, and hence imply that problem (1.1), (1.2) has at most one solution. In the latter case, using the method given by the author in [8, Lemma 2], we can obtain our Theorem 3. Here we give the proof as follows.

Proof of Theorem 3. Suppose conversely that $x_1(t), x_2(t)$ are different positive solutions of (1.1), (1.2), with

$$x_1(t_i) = x_2(t_i), \quad i = 1,2, \quad 0 < x_1(t) < x_2(t) \quad \text{for } t \in (t_1,t_2), \tag{4.1}$$

and at least one of them is a $C^1[0,1]$ solution. Here t_1, t_2 are some points in $[0,1]$ with $t_1 < t_2$. Then it must be true that $x_1(t) \in C^1[t_1,t_2]$, otherwise we would have $x_2(t) \in C^1[t_1,t_2]$ and $x_1'(t_1^+) = +\infty$ (or $x_1(t_2^-) = -\infty$) since $x_1''(t) \leq 0$, which is impossible because of (4.1).

Since $x_2''(t) \leq 0$ on $(t_1,t_2), x_2(t_1^+)(x_2(t_2^-))$ is either existent or $= +\infty(-\infty)$. This, together with (4.1), yields

$$\lim_{t\to t_1}(x_2'(t) - x_1'(t)) \geq 0, \qquad \lim_{t\to t_2}(x_2'(t) - x_1'(t)) \leq 0.$$

Let $y(t) = x_1(t)x_2'(t) - x_2(t)x_1'(t), t \in (t_1, t_2)$. The above discussion shows $y(t_1^+) \geq y(t_2^-)$. On the other hand, for $t \in (t_1, t_2)$,

$$\begin{aligned} y'(t) &= x_1(t)x_1''(t) - x_2(t)x_1''(t) \\ &= x_1(t)x_2(t)(f(t, x_1(t))/x_1(t) - f(t, x_2(t))/x_2(t)) > 0, \end{aligned}$$

which implies $y(t_2^-) > y(t_1^+)$, a contradiction that proves our conclusion. The proof is complete. □

Remark 1. Condition (U) in Theorem 3 is sharp in the sense that, when $f(t, x)$ is linear in x, the solution of (1.1), (1.2), if any, is not unique.

Remark 2. Besides [8], the function $y(t)$ has also been used to obtain a uniqueness result in [6].

Combining Theorems 2 and 3, we immediately get the following corollary.

COROLLARY 1. *Suppose that conditions* (E) *and* (U) *hold. Then problem* (1.1), (1.2) *has a unique positive solution which is a* $C^1[0, 1]$ *solution, provided* (1.6) *is fulfilled.*

5. Application to subcritical elliptic problems. To show the application of our results to elliptic boundary value problems we consider two kinds of problems.

First, we consider the Dirichlet problem of elliptic equations

$$\Delta u + \sum_{k=1}^{m} p_k(|x|)u^{\lambda_k} = 0, \qquad x \in \Omega,$$

(5.1)

$$u(x)|_{\partial\Omega} = 0, \qquad \left.\frac{\partial u}{\partial \vec{n}}\right|_{\partial\Omega} \text{ exists.}$$

Here Ω is an n-ball of radius ρ centered at $0 \in R^n, 0 < \rho < +\infty, n \geq 2, p_k(r) \in C[0, \rho), p_k > 0$ on $[0, \rho), -\infty < \lambda_k < 1, k = 1, 2, \ldots, m$.

As we show at the beginning of §1 (where we should note that some important relations have been omitted, among them $t'(r) = t^2/p(r), p(r)u'(r) = t\dot{y}(t) - y(t)$, and, when $y(t)$ is a $C^1[0, 1]$ solution of the final problem, $\lim_{t\to 0} y(t)/t = \dot{y}(0) > 0$ and $\lim_{r\to 0}(t\dot{y}(t) - y(t))/p(r) = \lim_{r\to 0} t\ddot{y}(t)t'(r)/p'(r) = -\lim_{r\to 0} p(r)f(r, y(t)/t)/p'(r) = 0$), seeking the positive radial solutions of (5.1) is equivalent to seeking the positive solutions of the following boundary value problem of ordinary differential equations:

$$\begin{aligned} &\ddot{y} + F(t, y) = 0, \qquad t \in (0, 1), \\ &y(0) = y(1) = 0, \qquad y(t) \in C^1[0, 1], \end{aligned}$$

where

$$F(t, y) = (r(t))^{2(n-1)} \sum_{k=1}^{m} (p_k(r(t))/t^{3+\lambda_k})y^{\lambda_k}.$$

Since

$$\int_0^1 F(t, t(1-t))\, dt = \int_0^\rho (r^{n-1}/t(r)) \sum_{k=1}^{m} p_k(r)(1 - t(r))^{\lambda_k}\, dr,$$

and

$$t(r) = \begin{cases} \left(\int_r^\rho s^{-(n-1)}\, ds + 1\right)^{-1} \sim r^{n-2}, & \text{as } r \to 0, \text{ if } n > 2 \\ (\ln\rho - \ln r + 1)^{-1}, & \text{if } n = 2, \end{cases}$$

$$(1 - t(r)) = \left(\int_r^\rho s^{-(n-1)}\, ds + 1\right)^{-1} \int_r^\rho s^{-(n-1)}\, ds \sim (\rho - r), \quad \text{as } r \to \rho,$$

the convergence of $\int_0^1 F(t,t(1-t))\,dt$ is equivalent to that of $\int_0^\rho r\sum_{k=1}^m p_k(r)(\rho-r)^{\lambda_k}\,dr$ (when $n>2$) or $\int_0^\rho r(\ln\rho-\ln r+1)\sum_{k=1}^m p_k(r)(\rho-r)^{\lambda_k}\,dr$ (when $n=2$). Hence, according to Theorem 2 and Corollary 1, and taking note of the continuity of $p_k(r)$ at $r=0$, we immediately deduce the following result.

COROLLARY 2. *Problem* (5.1) *possesses a positive radial solution if and only if* $\int_0^\rho p_k(r)(\rho-r)^{\lambda_k}\,dr<\infty$ *for each* $k=1,2,\dots,m$. *Moreover, the solution, if any, is unique.*

Now we consider the second problem

$$\begin{aligned}&\Delta u+f(|x|,u)=0,\quad x\in R^n,\quad n>2,\\ &\lim_{|x|\to\infty}u(x)=0.\end{aligned}\tag{5.2}$$

We give a nonexistence result as follows.

COROLLARY 3. *Suppose that* $f(r,u)=[0,\infty)\times(0,\infty)\to[0,\infty)$ *is continuous and satisfies condition* (E) *by replacing* t *there by* $r\in[0,\infty)$. *Then, if* $\int_0^\infty rf(r,1)\,dr$ *is nonconvergent, problem* (5.2) *admits no positive radial solutions.*

Proof. If (5.2) admits a positive radial solution, then, using the same transformation given at the beginning of §1, we will deduce that there exists at least a $C[0,1]$ positive solution (belonging to $C^1[0,1)$ in fact) of the problem

$$\begin{aligned}&\ddot{y}+(p^2(r(t))/t^3)f(r(t),y/t)=0\qquad t\in(0,1),\\ &y(0)=y(1)=0.\end{aligned}\tag{5.3}$$

But we have, on the other hand, that

$$\begin{aligned}\int_0^1 t(1-t)(p^2(r(t))/t^3)f(r(t),1/t)\,dt&=\int_0^\infty(1-t(r))p(r)f(r,1/t(r))\,dr\\ &\geq c\int_0^\infty(1-t(r))p(r)t(r)^{-\lambda}f(r,1)\,dr=\infty,\end{aligned}$$

since (E) holds and $t(r)\to 1, p(r)(1-t(r))\sim r$, as $r\to\infty$. Here $c>0$ is some constant. According to Theorem 1, the above relation implies that (5.3) cannot have a positive solution. This contradiction assures us that the conclusion of this corollary must be true. The proof is complete.

Remark. We refer to [10] and [11] for further information on Dirichlet problems concerning semilinear elliptic equations in bounded and unbounded domains. About the significance of finding radial solutions we recall the celebrated result in [12], which asserts that a solution of $\Delta u+f(u)=0, u(x)|_{\partial\Omega}=0$, where Ω is an n-ball, is indeed radially symmetric.

6. Appendix. Although the geometric significance of Lemma 2 is very clear, the analytic demonstration is somewhat complicated. The proof is divided into two steps.

Step 1. Let $g(t)\in C[0,1]$ be given such that $g(t)\geq 0$ on $[0,1]$ and $g(0)=g(1)=0$. We prove there exist point sequences (t_n,x_n) and $(\bar{t}_n,\bar{x}_n), n=0,1,2,\dots$, with $t_0=\bar{t}_0=\frac{1}{2}$ and $x_0=\bar{x}_0>0$, such that

1°. $0<\cdots<t_{n+1}<t_n<\cdots<t_1<t_0(=\frac{1}{2}), t_n\to 0$ as $n\to\infty$;

2°. $x_n\to 0$ as $n\to\infty, k_{n+1}>k_n>0$, and $k_n(t-t_n)+x_n\geq g(t)$ for $t\in[t_{n+1},t_n], n=0,1,2,\dots$, here $k_n=(x_{n+1}-x_n)/(t_{n+1}-t_n)$;

3°. $\bar{t}_0(=t_0) < \bar{t}_1 < \cdots < \bar{t}_n < \bar{t}_{n+1} < \cdots < 1, \bar{t}_n \to 1$ as $n \to \infty$;

4°. $\bar{x}_n \to 0$ as $n \to \infty, \bar{k}_{n+1} < \bar{k}_n < 0$ and $\bar{k}_n(t - \bar{t}_n) + \bar{x}_n \geq g(t)$ for $t \in [\bar{t}_n, \bar{t}_{n+1}], n = 0, 1, 2, \ldots$, here $\bar{k}_n = (\bar{x}_{n+1} - \bar{x}_n)/(\bar{t}_{n+1} - \bar{t}_n)$.

The existence of the points $(t_n, x_n), n = 0, 1, 2, \ldots$, for which 1° and 2° hold can be clarified as follows.

In fact, if there is an $I > 0$ such that $It \geq g(t)$ for $t \in [0, 1]$, then there is a $c_0 > 0$ such that, when $c > c_0, p(t) =: c[\frac{1}{4} - (t - \frac{1}{2})^2]^{1/2} \geq It \geq g(t)$ for $t \in [0, \frac{1}{2}]$. Fix $c > c_0$ sufficiently large and let $\{t_n\}$ be a sequence such that 1° holds. Take $x_n = p(t_n), n = 0, 1, 2, \ldots$; then 2° holds automatically since (t_n, x_n) are, successively, on an elliptic curve.

If the number I does not exist, then fix $t_0 = \frac{1}{2}, x_0 \geq 1 + \max g(t)$ at first, and then choose $(t_n, x_n), n = 1, 2, \ldots$, step by step as follows.

Let k_0 be the number such that $I_0(t) =: k_0(t - t_0) + x_0 \geq g(t)$ for $t \in [0, 1]$ and $I_0(\hat{t}) = g(\hat{t})$ for some $\hat{t} \in [0, t_0)$ (such a number is of course uniquely existent). It is easy to see that $k_0 > 0$ and there is a $t_0^* \in (0, t_0)$ such that $I_0(t) > g(t)$ for $t \in [0, t_0^*)$ and $I_0(t_0^*) = g(t_0^*)$.

Take $t_1 = \frac{1}{2}t_0^*, x_1 = I_0(t_1)$. We have $0 < x_1 < x_0$ and $(x_1 - x_0)/(t_1 - t_0) = k_0$. Denote by k_1 the number that satisfies $I_1(t) =: k_1(t - t_1) + x_1 \geq g(t)$ for $t \in [0, 1]$, and $I_1(\hat{t}) = g(\hat{t})$ for some $\hat{t} \in [0, t_1)$. We have, obviously, that $k_1 > k_0$ and there is a $t_1^* \in (0, t_1)$ such that $I_1(t) > g(t)$ for $t \in [0, t_1^*)$ and $I_1(t_1^*) = g(t_1^*)$.

We then take $t_2 = \frac{1}{2}t_1^*, x_2 = I_1(t_2)$. It implies that $0 < x_2 < x_1$ and $(x_2 - x_1)/(t_2 - t_1) = k_1$.

Generally, when $t_n = \frac{1}{2}t_{n-1}^*, x_n = I_{n-1}(t_n)(>0)$ have been taken, we denote by k_n the number that satisfies $I_n(t) =: k_n(t - t_n) + x_n \geq g(t)$ for $t \in [0, 1]$ and $I_n(\hat{t}) = g(\hat{t})$ for some $\hat{t} \in [0, t_n)$. Then it is true that $k_n > k_{n-1} = (x_n - x_{n-1})/(t_n - t_{n-1})$, and there is a $t_n^* \in (0, t_n)$ such that $I_n(t) > g(t)$ for $t \in [0, t_n^*), I_n(t_n^*) = g(t_n^*)$. We then take $t_{n+1} = t_n^*/2$ and $x_{n+1} = I_n(t_{n+1})$. Obviously, $0 < x_{n+1} < x_n$ and $(x_{n+1} - x_n)/(t_{n+1} - t_n) = k_n$.

So we have proved the existence of the sequence $(t_n, x_n), n = 0, 1, 2, \ldots$, which satisfies 1° and 2°. The points $(\bar{t}_n, \bar{x}_n), n = 0, 1, 2, \ldots$, for which 3° and 4° hold can be gotten in a similar way. And one can adjust x_0 or $\bar{x}_0$ so that $x_0 = \bar{x}_0$ holds. We leave the details to the reader.

Step 2. For any given constants a, b, δ_1, and δ_2 such that $\delta_1 > b - a > \delta_2$, let $m(t) = m(a, b, \delta_1, \delta_2, t) \in C^2[0, 1]$ be a function such that the following conditions hold:

$$\begin{aligned} &m(0) = a, \quad m(1) = b, \quad \dot{m}(0) = \delta_1, \quad \dot{m}(1) = \delta_2, \\ &\ddot{m}(0) = \ddot{m}(1) = 0 \quad \text{and} \quad \ddot{m}(t) \leq 0 \quad \text{for } t \in [0, 1]. \end{aligned} \tag{4.1}$$

To devise a way to construct such a function, let us consider the following example. Take an integer n such that $(n + 1)[\delta_1 - (b - a)] - [(b - a) - \delta_2] > 0$; write

$$\varepsilon = [(n + 1)(\delta_1 - (b - a)) - ((b - a) - \delta_2)]/[(b - a) - \delta_2](> 0)$$

and

$$I = \frac{\delta_1 - \delta_2}{n!}(\varepsilon + 1)(\varepsilon + 2)\ldots(\varepsilon + n + 1),$$

and let $m(t) = a + \delta_1 t - I\int_0^t (t - s)s^n(1 - s)^\varepsilon \, ds$. Then one can examine whether the conditions in (4.1) are all valid (to check this, one should notice that, for any integer n and $\varepsilon > 0, \int_0^1 s^n(1 - s)^\varepsilon \, ds = \int_0^1 (1 - s)^n s^\varepsilon \, ds = n!/(\varepsilon + 1)(\varepsilon + 2)\ldots(\varepsilon + n + 1)$).

Write

$$\delta_{1n} = \frac{k_n + k_{n+1}}{2}(t_n - t_{n+1}), \quad \delta_{2n} = \frac{k_n + k_{n-1}}{2}(t_n - t_{n+1}),$$
$$\bar{\delta}_{1n} = \frac{\bar{k}_n + \bar{k}_{n-1}}{2}(\bar{t}_{n+1} - \bar{t}_n) \quad \text{and} \quad \bar{\delta}_{2n} = \frac{\bar{k}_n + \bar{k}_{n+1}}{2}(\bar{t}_{n+1} - \bar{t}_n),$$

and let

$$m_n(t) = m(x_{n+1}, x_n, \delta_{1n}, \delta_{2n}, t), \bar{m}_n(t) = m(\bar{x}_n, \bar{x}_{n+1}, \bar{\delta}_{1n}, \bar{\delta}_{2n}, t),$$
$$t \in [0,1], n = 0, 1, 2, \ldots,$$

where t_n, x_n, k_n and $\bar{t}_n, \bar{x}_n, \bar{k}_n$ are those constants obtained in step 1, $k_{-1} = \bar{k}_0, \bar{k}_{-1} = k_0$. Now we define $q(t)$ as follows:

$$q(t) = \begin{cases} m_n((t - t_{n+1})/(t_n - t_{n+1})), & t \in [t_{n+1}, t_n], \\ m_n((t - \bar{t}_n)/(\bar{t}_{n+1} - \bar{t}_n)), & t \in [\bar{t}_n, \bar{t}_{n+1}], \\ 0 & t = 0, 1. \end{cases}$$

It is easy to check that $q(t) \in C[0,1] \cap C^2(0,1)$ and satisfies $q''(t) \leq 0$ for $t \in (0,1), q(t_n) = x_n, q(\bar{t}_n) = \bar{x}_n, n = 0, 1, 2, \ldots$. Also, $q(t) \geq k_n(t - t_n) + x_n \geq g(t)$ for $t \in [t_{n+1}, t_n]$ and $q(t) \geq \bar{k}_n(t - \bar{t}_n) + \bar{x}_n \geq g(t)$ for $t \in [\bar{t}_n, \bar{t}_{n+1}], n = 0, 1, 2, \ldots$, since $q(t)$ is a concave function. This completes the proof. □

REFERENCES

[1] S. D. TALIAFERRO, *A nonlinear singular boundary value problem*, Nonlinear Anal., 3 (1979), pp. 897–904.

[2] L. E. BOBISUD, D. O'REGAN, AND W. D. ROYALTY, *Singular boundary value problems*, Appl. Anal., 23 (1986), pp. 233–243.

[3] D. O'REGAN, *Positive solutions to singular and nonsingular second-order boundary value problems*, J. Math. Anal. Appl., 142 (1989), pp. 40–52.

[4] J. A. GATICA, V. OLIKER, AND P. WALTMAN, *Singular nonlinear boundary value problems for second-order ordinary differential equations*, J. Differential Equations, 79 (1989), pp. 62–78.

[5] A. M. FINK, J. A. GATICA, G. E. HERNANDEZ, AND P. WALTMAN, *Approximation of solutions of singular second-order boundary value problems*, SIAM J. Math. Anal., 22 (1991), pp. 440–462.

[6] C. D. LUNING AND W. L. PERRY, *Positive solutions of negative exponent generalized Emden–Fowler boundary value problem*, SIAM J. Math. Anal., 12 (1981), pp. 874–879.

[7] R. C. FLAGG, C. D. LUNING, AND W. L. PERRY, *Implementation of new iterative techniques for solutions of Thomas–Fermi and Emden–Fowler equations*, J. Comp. Phys., 38 (1980), pp. 396–405.

[8] Y. ZHANG, *Positive solutions of singular sublinear Emden–Fowler boundary value problems*, J. Math. Anal. Appl., 185 (1994), pp. 215–222.

[9] ———, *Existence of solutions of a kind of singular boundary value problems*, Nonlinear Anal., 21 (1993), pp. 153–159.

[10] J. SMOLLER AND A. WASSERMAN, *Existence of positive solutions for semilinear elliptic equations in general domains*, Arch. Rational Mech. Anal., 93 (1987), pp. 229–249.

[11] E. S. NOUSSAIR AND C. A. SWANSON, *An $L^q(R^n)$-theory of subcritical semilinear elliptic problems*, J. Differential Equations, 84 (1990), pp. 52–61.

[12] B. GIDAS, W. M. NI, AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[13] S. R. BERNFELD AND V. LAKSHMIKANTHAM, *An Introduction to Nonlinear Boundary Value Problems*, Math. in Sci. and Engrg., Vol. 109, Academic Press, New York, 1974.

[14] P. HARTMAN, *Ordinary Differential Equations*, 2nd Ed., Birkhauser, Boston, 1982.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 340–363, March 1995

005

PARAMETER DEPENDENCE OF PROPAGATION SPEED OF TRAVELLING WAVES FOR COMPETITION-DIFFUSION EQUATIONS*

YUKIO KAN-ON†

Abstract. In this paper, travelling wave solutions for certain competition-diffusion equations are considered, and the monotone dependence of their propagation speed on parameters which appear in the equation are established. To do this, the maximum principle and the bifurcation theory for heteroclinic orbits are employed.

Key words. travelling wave solution, competition-diffusion equation, exponential dichotomy

AMS subject classification. 34C37

1. Introduction.

There have been many studies of reaction-diffusion equations of the form

$$\tilde{u}_t = D\,\tilde{u}_{xx} + \tilde{f}(\tilde{u}), \quad x \in \mathbf{R}, \quad t > 0 \tag{1.1}$$

to explain phenomena that appear in various fields, where $\tilde{u} = (\tilde{u}_j)$ and $\tilde{f} = (\tilde{f}_j)$ are n-dimensional vectors and D is a diagonal constant $n \times n$ matrix. One interesting phenomenon is the appearance of travelling wave solutions which are of form $\tilde{U}(\xi)$, $\xi = x - s\,t$, where s is the propagation speed. To understand this phenomenon, we may study the properties of the solutions $(\tilde{U}, s)$, which satisfy the ODE

$$0 = D\,\tilde{U}_{\xi\xi} + s\,\tilde{U}_{\xi} + \tilde{f}(\tilde{U}), \quad \xi \in \mathbf{R}. \tag{1.2}$$

Comparatively speaking, when $n = 1$ we can easily study the existence of solutions of (1.2) by the analysis of motions in the phase plane, because the so-called *comparison principle* holds in the case of $n = 1$. We introduce here the following scalar equation as a typical and suggestive example:

$$u_t = u_{xx} + u\,(1-u)\,(u-\mu), \quad x \in \mathbf{R}, \quad t > 0. \tag{1.3}$$

For each $0 < \mu < 1$, it follows that (1.3) has two locally stable equilibria $u = 0$ and $u = 1$ in the ODE sense, and that

$$\begin{cases} 0 = u_{\xi\xi} + s\,u_{\xi} + u\,(1-u)\,(u-\mu), & \xi \in \mathbf{R}, \\ u(-\infty) = 0, \quad u(+\infty) = 1 \end{cases}$$

has a unique solution for $s = (2\,\mu - 1)/\sqrt{2}$ (for example, see Murray [11, pp. 304–305]). We note that the above propagation speed is monotone with respect to μ.

In general, the comparison principle does not always hold in case of $n \geq 2$. This leads to considerable complexity for studying travelling wave solutions of (1.1). In this paper, we discuss travelling wave solutions of the following competition-diffusion

* Received by the editors February 22, 1993; accepted for publication (in revised form) October 15, 1993.

† Department of Mathematics, Faculty of Education, Ehime University, Matsyama 790, Japan.

equation, where the comparison principle holds in the framework of (1.1) for $n \geq 2$:

$$\text{(1.4)} \qquad \begin{cases} u_t = u_{xx} + f(u,v), \\ v_t = d\,v_{xx} + g(u,v), \quad x \in \mathbf{R}, \quad t > 0, \\ u(0,x) \geq 0, \quad v(0,x) \geq 0, \quad x \in \mathbf{R}, \end{cases}$$

where

$$f(u,v) = u\,(1 - u - c\,v), \quad g(u,v) = v\,(a - b\,u - v),$$

and a, b, c, and d are positive constants which have ecological meanings.

When both u and v are spatially homogeneous, the evolution of (u,v) is governed by

$$u_t = f(u,v), \quad v_t = g(u,v), \quad t > 0.$$

We easily see that the asymptotic behavior of (u,v) with the initial conditions $u(0) > 0$ and $v(0) > 0$ is classified into the following four cases:

(I) If $a \leq \min(b, 1/c)$, then $\lim_{t \to +\infty}(u,v)(t) = (1,0)$.

(II) If $b < a < 1/c$, then

$$\lim_{t \to +\infty}(u,v)(t) = \left(\frac{1 - a\,c}{1 - b\,c}, \frac{a - b}{1 - b\,c}\right).$$

(III) If $1/c < a < b$, then $(0,a)$ and $(1,0)$ are locally stable and almost every solution converges to one of them as $t \to +\infty$.

(IV) If $a \geq \max(b, 1/c)$, then $\lim_{t \to +\infty}(u,v)(t) = (0,a)$.

In consideration of the situation for (1.3) with $0 < \mu < 1$, when (a,b,c) satisfies case (III), we guess that (1.4) has travelling wave solutions which decay exponentially to two locally stable equilibria, $(0,a)$ and $(1,0)$, and that their propagation speed is monotone with respect to the parameters a, b, and c. To justify our expectation, we study solutions of

$$\text{(1.5a)} \qquad \begin{cases} 0 = u_{\xi\xi} + s\,u_\xi + f(u,v), \\ 0 = d\,v_{\xi\xi} + s\,v_\xi + g(u,v), \quad \xi \in \mathbf{R} \end{cases}$$

with the boundary conditions

$$\text{(1.5b)} \qquad (u,v)(-\infty) = (0,a), \quad (u,v)(+\infty) = (1,0),$$

when (a,b,c) satisfies case (III).

There are many interesting studies for travelling wave solutions of (1.1). Vol'pert and Vol'pert [13] established a unique existence theorem of travelling wave solutions for (1.1) with the so-called *cooperative interaction process*, that is,

$$\frac{\partial \widetilde{f}_i}{\partial \widetilde{u}_j} > 0, \quad i,j = 1,2,\ldots,n, \quad i \neq j.$$

We note that (1.1) with the cooperative interaction process is one of the equations such that the comparison principle holds. On the other hand, for generalized competition-diffusion equations including (1.4), Gardner [4] and Conley and Gardner [2] proved the existence of travelling wave solutions by the topological method, so that their propagation speed is determined implicitly. (We also refer to Tang and Fife [12], Hosono and Mimura [6], and Hosono [5].) For the propagation speed of (1.4), Mimura and Fife [10] proved the existence of solutions with $s = 0$, and Ikeda and Mimura [7] showed numerically the monotone dependence on some parameters. Unfortunately,

we have not known rigorous results on the parameter dependence of the propagation speed. Our main purpose in this paper is to establish the monotone dependence of the propagation speed on the parameters a, b, and c.

Let us explain our approach briefly. Let $\boldsymbol{u}(\xi) = (u, v)(\xi)$ be a positive solution of (1.5) for some s. By formally differentiating (1.5a) with respect to $\mu \in \{a, b, c\}$, we obtain

$$\mathcal{L}\,\boldsymbol{u}_\mu(\xi) = -\begin{pmatrix} s_\mu\, u_\xi(\xi) + f_\mu(\boldsymbol{u}(\xi)) \\ s_\mu\, v_\xi(\xi) + g_\mu(\boldsymbol{u}(\xi)) \end{pmatrix}, \quad \xi \in \mathbf{R}, \tag{1.6}$$

where

$$\mathcal{L}(U, V) = \begin{pmatrix} U_{\xi\xi} + s\,U_\xi + f_u(\boldsymbol{u}(\xi))\,U + f_v(\boldsymbol{u}(\xi))\,V \\ d\,V_{\xi\xi} + s\,V_\xi + g_u(\boldsymbol{u}(\xi))\,U + g_v(\boldsymbol{u}(\xi))\,V \end{pmatrix}.$$

It is suggested from Lemma A.2 in Kan-on and Yanagida [8] that $\mathcal{L}^*(u, v) = 0$ has a nontrivial bounded solution $(u^*, v^*)(\xi)$, which satisfies $u^*(\xi)\,v^*(\xi) < 0$ for any $\xi \in \mathbf{R}$, where $\mathcal{L}^*$ is the adjoint operator of $\mathcal{L}$. By taking the inner product between (1.6) and $(u^*, v^*)(\xi)$, we obtain

$$\begin{aligned} 0 = & s_\mu \int_{\mathbf{R}} \{\, u_\xi(\xi)\,u^*(\xi) + v_\xi(\xi)\,v^*(\xi)\,\}\,d\xi \\ & + \int_{\mathbf{R}} \{\, f_\mu(\boldsymbol{u}(\xi))\,u^*(\xi) + g_\mu(\boldsymbol{u}(\xi))\,v^*(\xi)\,\}\,d\xi. \end{aligned}$$

If $\boldsymbol{u}(\xi)$ satisfies $u_\xi(\xi) > 0 > v_\xi(\xi)$ for any $\xi \in \mathbf{R}$, then we have

$$\begin{aligned} \frac{\partial s}{\partial a} &= -\frac{\int_{\mathbf{R}} v(\xi)\,v^*(\xi)\,d\xi}{\int_{\mathbf{R}} \{\, u_\xi(\xi)\,u^*(\xi) + v_\xi(\xi)\,v^*(\xi)\,\}\,d\xi} > 0, \\ \frac{\partial s}{\partial b} &= \frac{\int_{\mathbf{R}} u(\xi)\,v(\xi)\,v^*(\xi)\,d\xi}{\int_{\mathbf{R}} \{\, u_\xi(\xi)\,u^*(\xi) + v_\xi(\xi)\,v^*(\xi)\,\}\,d\xi} < 0, \\ \frac{\partial s}{\partial c} &= \frac{\int_{\mathbf{R}} u(\xi)\,v(\xi)\,u^*(\xi)\,d\xi}{\int_{\mathbf{R}} \{\, u_\xi(\xi)\,u^*(\xi) + v_\xi(\xi)\,v^*(\xi)\,\}\,d\xi} > 0. \end{aligned}$$

These inequalities suggest that the propagation speed is monotone with respect to a, b, and c. In order to justify the above argument, we shall employ the maximum principle and the bifurcation theory for heteroclinic orbits.

2. Statement of result. We shall say that $(u, v)(\xi)$ is (*strictly*) *monotone* if $u(\xi)$ is (strictly) increasing and $v(\xi)$ (strictly) decreasing. We state the following main theorem for any fixed $d > 0$.

THEOREM 2.1. *There exist families* $(u, v)(\xi; a, b, c)$ *and* $s(a, b, c)$ *defined on* $\mathcal{P} = \{\, (a, b, c) \mid 0 < 1/c < a < b \,\}$ *such that*

(i) $(u, v)(\xi; a, b, c)$ *is a strictly monotone solution of* (1.5) *with* $s = s(a, b, c)$ *for each* $(a, b, c) \in \mathcal{P}$,

(ii) $u(.; a, b, c)$, $v(.; a, b, c)$, *and* $s(a, b, c)$ *are a* C^1 *class in* (a, b, c), *and*

(iii) $s(a, b, c)$ *satisfies* $-2 < s(a, b, c) < 2\sqrt{a\,d}$,

$$\frac{\partial}{\partial a} s(a, b, c) > 0, \quad \frac{\partial}{\partial b} s(a, b, c) < 0, \quad \frac{\partial}{\partial c} s(a, b, c) > 0$$

for any $(a, b, c) \in \mathcal{P}$.

Furthermore, if $(\widetilde{u},\widetilde{v})(\xi)$ *is an arbitrary positive solution of* (1.5) *with* $s=\widetilde{s}$ *for* $(a,b,c)\in\mathcal{P}$, *then* $\widetilde{s}=s(a,b,c)$ *holds and there exists* η *such that* $(\widetilde{u},\widetilde{v})(\xi)=(u,v)(\xi+\eta;a,b,c)$ *for any* $\xi\in\mathbf{R}$.

We shall prove the above theorem in the following section.

3. Proofs.

LEMMA 3.1 (Theorem A.2 in Mimura and Fife [10]). *Suppose that* β *satisfies*

$$1<\beta<\min\left\{\frac{d+\sqrt{d^2+4}}{2},\frac{1+\sqrt{1+4d^2}}{2d}\right\}. \tag{3.1}$$

Then there exists $a_0\in(1/\beta,\beta)$ *such that* (1.5) *has a strictly monotone solution for* $(a,b,c,s)=(a_0,\beta,\beta,0)$.

Let $d>0$ be an arbitrary fixed constant. We denote by $\mathcal{E}$ the set of all parameters $(a,b,c)\in\mathcal{P}$ such that (1.5) has a strictly monotone solution for some s. Then we have $\mathcal{E}\neq\emptyset$ by virtue of Lemma 3.1.

Let $\boldsymbol{u}_0(\xi)=(u_0,v_0)(\xi)$ and s_0 be a strictly monotone solution and its propagation speed, respectively, of (1.5) for $(a,b,c)=(a_0,b_0,c_0)\in\mathcal{E}$. By the maximum principle, we have $\boldsymbol{u}_0(\xi)\in(0,1)\times(0,a_0)$ for any $\xi\in\mathbf{R}$. The linearized operators of (1.5a) around $(u,v)=(0,a_0)$ and $(u,v)=(1,0)$ are represented as

$$\mathcal{L}_0^-(u,v;s)=\begin{pmatrix} p_u^-(\frac{d}{d\xi};s)\,u \\ -a_0\,b_0\,u+p_v^-(\frac{d}{d\xi};s)\,v\end{pmatrix}$$

and

$$\mathcal{L}_0^+(u,v;s)=\begin{pmatrix} p_u^+(\frac{d}{d\xi};s)\,u-c_0\,v \\ p_v^+(\frac{d}{d\xi};s)\,v\end{pmatrix},$$

respectively, where

$$p_u^-(\gamma;s)=\gamma^2+s\gamma+1-a_0\,c_0,\quad p_v^-(\gamma;s)=d\gamma^2+s\gamma-a_0,$$
$$p_u^+(\gamma;s)=\gamma^2+s\gamma-1,\quad p_v^+(\gamma;s)=d\gamma^2+s\gamma+a_0-b_0.$$

We note that $1-a_0\,c_0<0$ and $a_0-b_0<0$ hold for any $(a_0,b_0,c_0)\in\mathcal{P}$. With $*=u,\,v$, we denote by $\lambda_*^\pm(s)$ $(\lambda_*^-(s)<0<\lambda_*^+(s))$ and $\sigma_*^\pm(s)$ $(\sigma_*^-(s)<0<\sigma_*^+(s))$ the solutions of the quadratic equations $p_*^-(\gamma;s)=0$ and $p_*^+(\gamma;s)=0$, respectively. And we put

$$\Lambda_1^\pm(s)=\min\{\lambda_u^\pm(s),\lambda_v^\pm(s)\},\quad \Lambda_2^\pm(s)=\max\{\lambda_u^\pm(s),\lambda_v^\pm(s)\},$$
$$\Sigma_1^\pm(s)=\min\{\sigma_u^\pm(s),\sigma_v^\pm(s)\},\quad \Sigma_2^\pm(s)=\max\{\sigma_u^\pm(s),\sigma_v^\pm(s)\}.$$

From $p_u^\pm(-s;s)<0$ and $p_v^\pm(-s/d;s)<0$, we have

$$\begin{aligned}&\Sigma_1^-(s)\le\sigma_u^-(s)<-s<\lambda_u^+(s)\le\Lambda_2^+(s),\\ &\Sigma_1^-(s)\le\sigma_v^-(s)<-s/d<\lambda_v^+(s)\le\Lambda_2^+(s)\end{aligned} \tag{3.2}$$

for any $s\in\mathbf{R}$ and $(a_0,b_0,c_0)\in\mathcal{P}$. With $*=u,\,v$, by differentiating $p_*^-(\lambda_*^+(s);s)=0$ and $p_*^+(\sigma_*^-(s);s)=0$ with respect to s, we obtain

$$\frac{\partial}{\partial s}\lambda_*^+(s)=-\lambda_*^+(s)/\frac{d}{d\gamma}p_*^-(\lambda_*^+(s);s)<0,$$
$$\frac{\partial}{\partial s}\sigma_*^-(s)=-\sigma_*^-(s)/\frac{d}{d\gamma}p_*^+(\sigma_*^-(s);s)<0$$

for any $s\in\mathbf{R}$ and $(a_0,b_0,c_0)\in\mathcal{P}$. Then we see that $\lambda_u^+(s)$, $\lambda_v^+(s)$, $\sigma_u^-(s)$, and $\sigma_v^-(s)$ are strictly decreasing with respect to s for each $(a_0,b_0,c_0)\in\mathcal{P}$. Furthermore, it

follows that

$$\overline{\Phi}_1^{\pm}(\xi;s)=\begin{cases}\left(\dfrac{p_v^-(\lambda_u^{\pm}(s);s)}{a_0\,b_0},1\right)e^{\Lambda_1^{\pm}(s)\,\xi} & \text{if } \lambda_u^{\pm}(s)<\lambda_v^{\pm}(s),\\ -\left(\dfrac{\frac{d}{d\gamma}p_v^-(\lambda_v^{\pm}(s);s)}{a_0\,b_0},\xi\right)e^{\Lambda_1^{\pm}(s)\,\xi} & \text{if } \lambda_u^{\pm}(s)=\lambda_v^{\pm}(s),\\ (0,1)\,e^{\Lambda_1^{\pm}(s)\,\xi} & \text{if } \lambda_u^{\pm}(s)>\lambda_v^{\pm}(s),\end{cases}$$

$$\overline{\Phi}_2^{\pm}(\xi;s)=\begin{cases}(0,1)\,e^{\Lambda_2^{\pm}(s)\,\xi} & \text{if } \lambda_u^{\pm}(s)\le\lambda_v^{\pm}(s),\\ \left(\dfrac{p_v^-(\lambda_u^{\pm}(s);s)}{a_0\,b_0},1\right)e^{\Lambda_2^{\pm}(s)\,\xi} & \text{if } \lambda_u^{\pm}(s)>\lambda_v^{\pm}(s)\end{cases}$$

and

$$\overline{\Psi}_1^{\pm}(\xi;s)=\begin{cases}(1,0)\,e^{\Sigma_1^{\pm}(s)\,\xi} & \text{if } \sigma_u^{\pm}(s)\le\sigma_v^{\pm}(s),\\ \left(1,\dfrac{p_u^+(\sigma_v^{\pm}(s);s)}{c_0}\right)e^{\Sigma_1^{\pm}(s)\,\xi} & \text{if } \sigma_u^{\pm}(s)>\sigma_v^{\pm}(s),\end{cases}$$

$$\overline{\Psi}_2^{\pm}(\xi;s)=\begin{cases}\left(1,\dfrac{p_u^+(\sigma_v^{\pm}(s);s)}{c_0}\right)e^{\Sigma_2^{\pm}(s)\,\xi} & \text{if } \sigma_u^{\pm}(s)<\sigma_v^{\pm}(s),\\ \left(\xi,\dfrac{\frac{d}{d\gamma}p_u^+(\sigma_u^{\pm}(s);s)}{c_0}\right)e^{\Sigma_2^{\pm}(s)\,\xi} & \text{if } \sigma_u^{\pm}(s)=\sigma_v^{\pm}(s),\\ (1,0)\,e^{\Sigma_2^{\pm}(s)\,\xi} & \text{if } \sigma_u^{\pm}(s)>\sigma_v^{\pm}(s)\end{cases}$$

are linearly independent solutions of $\mathcal{L}_0^-(u,v;s)=0$ and $\mathcal{L}_0^+(u,v;s)=0$, respectively. By the above behavior, for any s, we have

$$\overline{\Phi}_2^{\pm}(\xi;s)=o(|\,\overline{\Phi}_1^{\pm}(\xi;s)\,|),\ \overline{\Phi}_1^{+}(\xi;s)=o(|\,\overline{\Phi}_2^{-}(\xi;s)\,|)\quad\text{as } \xi\to-\infty,$$
$$\overline{\Psi}_1^{\pm}(\xi;s)=o(|\,\overline{\Psi}_2^{\pm}(\xi;s)\,|),\ \overline{\Psi}_2^{-}(\xi;s)=o(|\,\overline{\Psi}_1^{+}(\xi;s)\,|)\quad\text{as } \xi\to+\infty.$$

Let us get the asymptotic expansion of $\boldsymbol{u}_0(\xi)$ as $\xi\to\pm\infty$. By $\boldsymbol{u}_0(-\infty)=(0,a_0)$, we see that $\boldsymbol{u}_0(\xi)$ satisfies

$$\begin{aligned}0&=u_{0\xi\xi}(\xi)+s_0\,u_{0\xi}(\xi)+f(\boldsymbol{u}_0(\xi))\\&=p_u^-(\tfrac{d}{d\xi};s_0)\,u_0(\xi)-u_0(\xi)\,\{\,u_0(\xi)+c_0\,(v_0(\xi)-a_0)\,\}\\&=p_u^-(\tfrac{d}{d\xi};s_0)\,u_0(\xi)+o(u_0(\xi)),\end{aligned}\tag{3.3}$$

$$\begin{aligned}0&=d\,v_{0\xi\xi}(\xi)+s_0\,v_{0\xi}(\xi)+g(\boldsymbol{u}_0(\xi))\\&=p_v^-(\tfrac{d}{d\xi};s_0)\,(v_0(\xi)-a_0)-a_0\,b_0u_0(\xi)\\&\qquad+(v_0(\xi)-a_0)\,(a_0-b_0\,u_0(\xi)-v_0(\xi))\\&=p_v^-(\tfrac{d}{d\xi};s_0)\,(v_0(\xi)-a_0)-a_0\,b_0u_0(\xi)+o(v_0(\xi)-a_0)\end{aligned}\tag{3.4}$$

as $\xi\to-\infty$. Since $0<u_0(\xi)<1$ holds for any $\xi\in\mathbf{R}$, it follows from (3.3) that $u_0(\xi)$ satisfies

$$u_0(\xi)=C_u^-\,e^{\lambda_u^+(s_0)\,\xi}\,(1+o(1))$$

as $\xi\to-\infty$, where C_u^- is a positive constant. By substituting the above expansion

into (3.4), we have

$$v_0(\xi) - a_0 = \begin{cases} \dfrac{a_0 b_0 C_u^- e^{\lambda_u^+(s_0)\xi}}{p_v^-(\lambda_u^+(s_0); s_0)}(1+o(1)) & \text{for } \lambda_u^+(s_0) < \lambda_v^+(s_0), \\ \dfrac{a_0 b_0 C_u^- \xi e^{\lambda_u^+(s_0)\xi}}{\frac{d}{d\gamma}p_v^-(\lambda_u^+(s_0); s_0)}(1+o(1)) & \text{for } \lambda_u^+(s_0) = \lambda_v^+(s_0), \\ \dfrac{a_0 b_0 C_u^- e^{\lambda_u^+(s_0)\xi}}{p_v^-(\lambda_u^+(s_0); s_0)}(1+o(1)) & \\ \quad +C_1 e^{\lambda_v^+(s_0)\xi}(1+o(1)) & \text{for } \lambda_u^+(s_0) > \lambda_v^+(s_0) \end{cases}$$

as $\xi \to -\infty$, where C_1 is some constant. Since $0 < v_0(\xi) < a_0$ holds for any $\xi \in \mathbf{R}$, and since $p_v^-(\lambda_u^+(s_0); s_0) > 0$ holds if $\lambda_u^+(s_0) > \lambda_v^+(s_0)$, we see that C_1 must satisfy $C_1 < 0$. By summarizing the above, we get the following asymptotic expansion:

$$\text{(3.5)} \qquad \begin{cases} u_0(\xi) = C_u^- e^{\lambda_u^+(s_0)\xi}(1+o(1)), \\ v_0(\xi) = a_0 - C_v^- |\xi|^{m_+} e^{\Lambda_1^+(s_0)\xi}(1+o(1)) \end{cases}$$

as $\xi \to -\infty$, where

$$m_\pm = \begin{cases} 1 & \text{if } \lambda_u^\pm(s_0) = \lambda_v^\pm(s_0), \\ 0 & \text{otherwise,} \end{cases}$$

and C_v^- is a positive constant which satisfies

$$C_v^- = \begin{cases} -\dfrac{a_0 b_0 C_u^-}{p_v^-(\lambda_u^+(s_0); s_0)} & \text{for } \lambda_u^+(s_0) < \lambda_v^+(s_0), \\ \dfrac{a_0 b_0 C_u^-}{\frac{d}{d\gamma}p_v^-(\lambda_u^+(s_0); s_0)} & \text{for } \lambda_u^+(s_0) = \lambda_v^+(s_0), \\ -C_1 & \text{for } \lambda_u^+(s_0) > \lambda_v^+(s_0), \end{cases}$$

because $p_v^-(\gamma; s)$ satisfies $p_v^-(\lambda_u^+(s); s) < 0$ if $\lambda_u^+(s) < \lambda_v^+(s)$, and $\frac{d}{d\xi}p_v^-(\lambda_u^+(s); s) > 0$ if $\lambda_u^+(s) = \lambda_v^+(s)$. Analogously we obtain

$$\text{(3.6)} \qquad \begin{cases} u_0(\xi) = 1 - C_u^+ |\xi|^{n_-} e^{\Sigma_2^-(s_0)\xi}(1+o(1)), \\ v_0(\xi) = C_v^+ e^{\sigma_v^-(s_0)\xi}(1+o(1)) \end{cases}$$

as $\xi \to +\infty$, where

$$n_\pm = \begin{cases} 1 & \text{if } \sigma_u^\pm(s_0) = \sigma_v^\pm(s_0), \\ 0 & \text{otherwise,} \end{cases}$$

and C_u^+, C_v^+ are some positive constants. Since $\boldsymbol{u}_0(\xi)$ is on both the unstable manifold at $(u, v) = (0, a_0)$ and the stable manifold at $(u, v) = (1, 0)$, we see from (3.5) and (3.6) that $\boldsymbol{u}_0(\xi)$ satisfies

$$\text{(3.7)} \qquad \boldsymbol{u}_0(\xi) = \begin{cases} (0, a_0) - C_v^- \overline{\Phi}_1^+(\xi; s_0) + o(|\overline{\Phi}_1^+(\xi; s_0)|) & \text{as } \xi \to -\infty, \\ (1, 0) - C_u^+ \overline{\Psi}_2^-(\xi; s_0) + o(|\overline{\Psi}_2^-(\xi; s_0)|) & \text{as } \xi \to +\infty. \end{cases}$$

3.1. Linearized operator. We define the linearized operator $\mathcal{L}$ of (1.5a) around $(u,v)=\boldsymbol{u}_0(\xi)$ and its adjoint operator $\mathcal{L}^*$ by

$$\mathcal{L}(u,v)=\begin{pmatrix} u_{\xi\xi}+s_0\,u_\xi+f_u^0(\xi)\,u+f_v^0(\xi)\,v \\ \\ d\,v_{\xi\xi}+s_0\,v_\xi+g_u^0(\xi)\,u+g_v^0(\xi)\,v \end{pmatrix}$$

and

$$\mathcal{L}^*(u,v)=\begin{pmatrix} u_{\xi\xi}-s_0\,u_\xi+f_u^0(\xi)\,u+g_u^0(\xi)\,v \\ \\ d\,v_{\xi\xi}-s_0\,v_\xi+f_v^0(\xi)\,u+g_v^0(\xi)\,v \end{pmatrix},$$

respectively, where $f_u^0(\xi)=f_u(\boldsymbol{u}_0(\xi))$ and the other functions f_v^0, g_u^0, and g_v^0 are also defined in the same manner with f_u^0.

Since $\boldsymbol{u}_0(\xi)$ is a strictly monotone solution of (1.5) with $s=s_0$, it follows that $\boldsymbol{u}_{0\xi}(\xi)$ is a nontrivial bounded solution of $\mathcal{L}(u,v)=0$ and satisfies $u_{0\xi}(\xi)\geq 0\geq v_{0\xi}(\xi)$ for any $\xi\in\mathbf{R}$. We assume that $\boldsymbol{u}_{0\xi}(\xi)$ satisfies $u_{0\xi}(\xi_1)=0$ and/or $v_{0\xi}(\xi_1)=0$ for some $\xi_1\in\mathbf{R}$. By $f_v^0(\xi)<0$ and $g_u^0(\xi)<0$ for any $\xi\in\mathbf{R}$, we have $\boldsymbol{u}_{0\xi}(\xi_1)=0$ and $\boldsymbol{u}_{0\xi\xi}(\xi_1)=0$. By uniqueness, we obtain $\boldsymbol{u}_{0\xi}(\xi)=0$ for any $\xi\in\mathbf{R}$, that is, $\boldsymbol{u}_0(\xi)$ is a constant function. This contradiction implies that $\boldsymbol{u}_0(\xi)$ satisfies $u_{0\xi}(\xi)>0>v_{0\xi}(\xi)$ for any $\xi\in\mathbf{R}$.

LEMMA 3.2. *Suppose that* $\boldsymbol{u}(\xi)=(u,v)(\xi)$ *is a nontrivial solution of* $\mathcal{L}(u,v)=0$, *which satisfies* $|\,\boldsymbol{u}(\xi)\,|=O(e^{\Lambda_2^+(s_0)\,\xi})$ *as* $\xi\to-\infty$ (*respectively,* $|\,\boldsymbol{u}(\xi)\,|=O(e^{\Sigma_1^-(s_0)\,\xi})$ *as* $\xi\to+\infty$) *and* $u(\xi)\,v(\xi)>0$ *for any* ξ *in a neighborhood of* $\xi=-\infty$ (*respectively,* $\xi=+\infty$). *Then* $\boldsymbol{u}(\xi)$ *satisfies* $u(\xi)\,v(\xi)>0$ *for any* $\xi\in\mathbf{R}$ *and* $\limsup_{\xi\to+\infty}|\,v(\xi)\,|>0$ (*respectively,* $\limsup_{\xi\to-\infty}|\,u(\xi)\,|>0$).

Proof. We show only the proof for the former case, because the latter can be proved in a similar manner. Furthermore, without loss of generality, we may assume $u(\xi)>0$ and $v(\xi)>0$ near $\xi=-\infty$.

We assume one of the following two cases: (i) $u(\xi)$ and/or $v(\xi)$ have zeros, and (ii) $\boldsymbol{u}(\xi)$ satisfies $u(\xi)\,v(\xi)>0$ for any $\xi\in\mathbf{R}$ and $\lim_{\xi\to+\infty}v(\xi)=0$. We put

$$\xi_2=\begin{cases}\inf\{\,\xi\,|\,u(\xi)\,v(\xi)\leq 0\,\} & \text{for the case (i),}\\ +\infty & \text{for the case (ii).}\end{cases}$$

Since $f_v^0(\xi)<0$ and $g_u^0(\xi)<0$ hold for any $\xi\in\mathbf{R}$, with the use of (3.2) and the integration by parts we have

$$\begin{aligned}0&=\int_{-\infty}^{\xi_2}\{\,u_{\xi\xi}(\xi)+s_0\,u_\xi(\xi)+f_u^0(\xi)\,u(\xi)+f_v^0(\xi)\,v(\xi)\,\}\,u_{0\xi}(\xi)\,e^{s_0\,\xi}\,d\xi\\ &=u_\xi(\xi_2)\,u_{0\xi}(\xi_2)\,e^{s_0\,\xi_2}\\ &\qquad+\int_{-\infty}^{\xi_2}f_v^0(\xi)\,\{\,v(\xi)\,u_{0\xi}(\xi)-u(\xi)\,v_{0\xi}(\xi)\,\}\,e^{s_0\,\xi}\,d\xi<0\end{aligned}$$

when $u(\xi_2)=0$, and

$$\begin{aligned}0&=\int_{-\infty}^{\xi_2}\{\,d\,v_{\xi\xi}(\xi)+s_0\,v_\xi(\xi)+g_u^0(\xi)\,u(\xi)+g_v^0(\xi)\,v(\xi)\,\}\,v_{0\xi}(\xi)\,e^{(s_0\,\xi)/d}\,d\xi\\ &=d\,v_\xi(\xi_2)\,v_{0\xi}(\xi_2)\,e^{(s_0\,\xi_2)/d}\\ &\qquad+\int_{-\infty}^{\xi_2}g_u^0(\xi)\,\{\,u(\xi)\,v_{0\xi}(\xi)-v(\xi)\,u_{0\xi}(\xi)\,\}\,e^{(s_0\,\xi)/d}\,d\xi>0\end{aligned}$$

when $v(\xi_2) = 0$ and/or $\xi_2 = +\infty$. These contradictions imply that the desired result holds. □

3.2. Fundamental solutions.

LEMMA 3.3 (Theorem 8.1 in Coddington and Levinson [1]). *Let $B(\xi)$ be an arbitrary 2×2 matrix which satisfies $B(\xi) = O(e^{-\gamma|\xi|})$ as $\xi \to -\infty$ (respectively, $\xi \to +\infty$) for some positive constant γ. Then there exists a fundamental set $\{\phi_1^{\pm}(\xi), \phi_2^{\pm}(\xi)\}$ of solutions of*

$$\mathcal{L}_0^-(\boldsymbol{u}; s) + B(\xi)\,\boldsymbol{u} = 0 \quad \big(\textit{respectively, } \mathcal{L}_0^+(\boldsymbol{u}; s) + B(\xi)\,\boldsymbol{u} = 0\big),$$

which satisfy

$$\phi_j^{\pm}(\xi) = \overline{\Phi}_j^{\pm}(\xi; s) + o(|\,\overline{\Phi}_j^{\pm}(\xi; s)\,|) \quad \textit{as } \xi \to -\infty,$$
$$\Big(\textit{respectively, } \phi_j^{\pm}(\xi) = \overline{\Psi}_j^{\pm}(\xi; s) + o(|\,\overline{\Psi}_j^{\pm}(\xi; s)\,|) \quad \textit{as } \xi \to +\infty\Big).$$

By (3.5) and (3.6), we have

$$\begin{pmatrix} f_u^0(\xi) & f_v^0(\xi) \\ g_u^0(\xi) & g_v^0(\xi) \end{pmatrix} = \begin{cases} \begin{pmatrix} 1 - a_0 c_0 & 0 \\ -a_0 b_0 & -a_0 \end{pmatrix} + o(e^{(\Lambda_1^+(s_0)-\delta)\xi}) \\ \qquad\qquad \text{as } \xi \to -\infty, \\ \begin{pmatrix} -1 & -c_0 \\ 0 & a_0 - b_0 \end{pmatrix} + o(e^{(\Sigma_2^-(s_0)+\delta)\xi}) \\ \qquad\qquad \text{as } \xi \to +\infty \end{cases}$$

for any $0 < \delta < \min\{\Lambda_1^+(s_0), -\Sigma_2^-(s_0)\}$. Then it follows from Lemma 3.3 that there exist fundamental sets $\{\Phi_1^{\pm}(\xi), \Phi_2^{\pm}(\xi)\}$ and $\{\Psi_1^{\pm}(\xi), \Psi_2^{\pm}(\xi)\}$ of solutions of $\mathcal{L}(u, v) = 0$ which satisfy

$$\Phi_j^{\pm}(\xi) = \overline{\Phi}_j^{\pm}(\xi; s_0) + o(|\,\overline{\Phi}_j^{\pm}(\xi; s_0)\,|) \quad \text{as } \xi \to -\infty,$$
$$\Psi_j^{\pm}(\xi) = \overline{\Psi}_j^{\pm}(\xi; s_0) + o(|\,\overline{\Psi}_j^{\pm}(\xi; s_0)\,|) \quad \text{as } \xi \to +\infty.$$

We also see that there exists a nonsingular constant 4×4 matrix $M = (M_{ij})$ such that

$$(\Phi_1^-, \Phi_2^-, \Phi_1^+, \Phi_2^+)(\xi) = (\Psi_1^-, \Psi_2^-, \Psi_1^+, \Psi_2^+)(\xi)\,M,$$
$$(\Psi_1^-, \Psi_2^-, \Psi_1^+, \Psi_2^+)(\xi) = (\Phi_1^-, \Phi_2^-, \Phi_1^+, \Phi_2^+)(\xi)\,\widetilde{M}$$

hold for any $\xi \in \mathbf{R}$, where $\widetilde{M} = (\widetilde{M}_{ij})$ is the inverse matrix of M. Furthermore, by the asymptotic behaviors of $\overline{\Phi}_j^{\pm}(\xi; s)$ and $\overline{\Psi}_j^{\pm}(\xi; s)$, we have

$$\Phi_2^{\pm}(\xi) = o(|\,\Phi_1^{\pm}(\xi)\,|),\ \Phi_1^+(\xi) = o(|\,\Phi_2^-(\xi)\,|) \quad \text{as } \xi \to -\infty,$$
$$\Psi_1^{\pm}(\xi) = o(|\,\Psi_2^{\pm}(\xi)\,|),\ \Psi_2^-(\xi) = o(|\,\Psi_1^+(\xi)\,|) \quad \text{as } \xi \to +\infty.$$

By the fact that an arbitrary solution $(u, v)(\xi)$ of $\mathcal{L}(u, v) = 0$ satisfies

$$\begin{aligned} 0 =& u_{\xi\xi}(\xi) + s_0\,u_\xi(\xi) + f_u^0(\xi)\,u(\xi) + f_v^0(\xi)\,v(\xi) \\ =& p_u^-\left(\tfrac{d}{d\xi}; s_0\right) u(\xi) - c_0\,u_0(\xi)\,v(\xi) + o(u(\xi)) \\ =& p_u^-\left(\tfrac{d}{d\xi}; s_0\right) u(\xi) - c_0\,C_u^-\,e^{\lambda_u^+(s_0)\,\xi}\,(1 + o(1))\,v(\xi) + o(u(\xi)) \end{aligned}$$

as $\xi \to -\infty$, we obtain

$$\Phi_2^{\pm}(\xi)\, e^{-\Lambda_2^{\pm}(s_0)\,\xi} = \begin{cases} \left(\dfrac{p_v^-(\lambda_u^{\pm}(s_0); s_0)}{a_0\, b_0},\ 1 \right) (1 + o(1)) \\ \qquad\qquad \text{if } \lambda_u^{\pm}(s_0) > \lambda_v^{\pm}(s_0), \\ \left(\dfrac{c_0\, C_u^-\, e^{\lambda_u^+(s_0)\,\xi}}{p_u^-(\lambda_u^+(s_0) + \lambda_v^{\pm}(s_0); s_0)},\ 1 \right) (1 + o(1)) \\ \qquad\qquad \text{if } \lambda_u^{\pm}(s_0) \le \lambda_v^{\pm}(s_0) \end{cases}$$

as $\xi \to -\infty$. Similarly, we also have

$$\Psi_1^{\pm}(\xi)\, e^{-\Sigma_1^{\pm}(s_0)\,\xi} = \begin{cases} \left(1,\ \dfrac{p_u^+(\sigma_v^{\pm}(s_0); s_0)}{c_0} \right) (1 + o(1)) \\ \qquad\qquad \text{if } \sigma_u^{\pm}(s_0) > \sigma_v^{\pm}(s_0), \\ \left(1,\ \dfrac{b_0\, C_v^+\, e^{\sigma_v^-(s_0)\,\xi}}{p_v^+(\sigma_u^{\pm}(s_0) + \sigma_v^-(s_0); s_0)} \right) (1 + o(1)) \\ \qquad\qquad \text{if } \sigma_u^{\pm}(s_0) \le \sigma_v^{\pm}(s_0) \end{cases}$$

as $\xi \to +\infty$.

We define the order relations $\succeq_s$ and $\succeq_o$ by the following manner:

$$(u_1, v_1) \succeq_s (u_2, v_2) \Longleftrightarrow u_1 \ge u_2 \text{ and } v_1 \ge v_2,$$
$$(u_1, v_1) \succeq_o (u_2, v_2) \Longleftrightarrow u_1 \ge u_2 \text{ and } v_1 \le v_2.$$

And we denote by $\succ_s$ and $\succ_o$ the relations which are defined by replacing $\le$ with $<$. Since

$$\lambda_u^-(s_0) \le \lambda_v^-(s_0) < \lambda_u^+(s_0) + \lambda_v^-(s_0) < \lambda_u^+(s_0)$$

holds when $\lambda_u^-(s_0) \le \lambda_v^-(s_0)$, we have

$$\begin{cases} p_v^-(\lambda_u^-(s_0); s_0) < 0 & \text{if } \lambda_u^-(s_0) > \lambda_v^-(s_0), \\ p_v^-(\lambda_u^+(s_0); s_0) > 0 & \text{if } \lambda_u^+(s_0) > \lambda_v^+(s_0), \\ p_u^-(\lambda_u^+(s_0) + \lambda_v^-(s_0); s_0) < 0 & \text{if } \lambda_u^-(s_0) \le \lambda_v^-(s_0), \\ p_u^-(\lambda_u^+(s_0) + \lambda_v^+(s_0); s_0) > 0 & \text{if } \lambda_u^+(s_0) \le \lambda_v^+(s_0), \end{cases} \tag{3.8}$$

that is, $\Phi_2^-(\xi) \prec_o (0,0)$ and $\Phi_2^+(\xi) \succ_s (0,0)$ as $\xi \to -\infty$. Similarly, by the fact that

$$\sigma_v^-(s_0) < \sigma_u^+(s_0) + \sigma_v^-(s_0) < \sigma_u^+(s_0) \le \sigma_v^+(s_0)$$

holds if $\sigma_u^+(s_0) \le \sigma_v^+(s_0)$, we also get $\Psi_1^-(\xi) \succ_s (0,0)$ and $\Psi_1^+(\xi) \succ_o (0,0)$ as $\xi \to +\infty$. From Lemma 3.2, we obtain

$$\Phi_2^+(\xi) \succ_s (0,0),\ \Psi_1^-(\xi) \succ_s (0,0) \quad \text{for any } \xi \in \mathbf{R},$$
$$\limsup_{\xi \to +\infty} [\Phi_2^+(\xi)]_2 > 0, \quad \limsup_{\xi \to -\infty} [\Psi_1^-(\xi)]_1 > 0,$$

where we denote by $[\boldsymbol{u}]_j$ the jth element of $\boldsymbol{u}$.

We assume $M_{44} = 0$, that is,

$$\Phi_2^+(\xi) = M_{14}\, \Psi_1^-(\xi) + M_{24}\, \Psi_2^-(\xi) + M_{34}\, \Psi_1^+(\xi).$$

If $M_{34} = 0$, then we have $\lim_{\xi\to+\infty} \Phi_2^+(\xi) = 0$. This contradiction implies that $M_{34} \neq 0$ holds. By $\lim_{\xi\to+\infty}[\Psi_1^+(\xi)]_1 = +\infty$, we get

$$M_{34}\,[\Phi_2^+(\xi)]_1 = M_{34}^2\,[\Psi_1^+(\xi)]_1\,(1+o(1)) > 0$$

as $\xi \to +\infty$. Since $\Phi_2^+(\xi) \succ_s (0,0)$ for any $\xi \in \mathbf{R}$, we have $M_{34} > 0$. By $[\Psi_1^+(\xi)]_2 < 0$ as $\xi \to +\infty$, we obtain

$$0 < \limsup_{\xi\to+\infty}[\Phi_2^+(\xi)]_2 = M_{34} \limsup_{\xi\to+\infty}[\Psi_1^+(\xi)]_1 \le 0.$$

This contradiction implies that $M_{44} \neq 0$ holds. Similarly, we also obtain $\widetilde{M}_{11} \neq 0$. Therefore we see that Φ_2^+ and $\Psi_1^-(\xi)$ satisfy

$$\begin{aligned}
\Phi_2^+(\xi) &= M_{44}\,\Psi_2^+(\xi) + o(|\,\Psi_2^+(\xi)\,|) \quad \text{as } \xi \to +\infty,\\
\Psi_1^-(\xi) &= \widetilde{M}_{11}\,\Phi_1^-(\xi) + o(|\,\Phi_1^-(\xi)\,|) \quad \text{as } \xi \to -\infty.
\end{aligned}$$

Since $\boldsymbol{u}_{0\xi}(\xi)$ is a bounded solution of $\mathcal{L}(u,v) = 0$, it follows that $\boldsymbol{u}_{0\xi}(\xi)$ satisfies

$$\boldsymbol{u}_{0\xi}(\xi) = \begin{cases} M_{30}\,\Phi_1^+(\xi) + M_{40}\,\Phi_2^+(\xi) \\ \qquad = M_{30}\,\overline{\Phi}_1^+(\xi;s_0) + o(|\,\overline{\Phi}_1^+(\xi;s_0)\,|) \quad \text{as } \xi \to -\infty, \\ M_{10}\,\Psi_1^-(\xi) + M_{20}\,\Psi_2^-(\xi) \\ \qquad = M_{20}\,\overline{\Psi}_2^-(\xi;s_0) + o(|\,\overline{\Psi}_2^-(\xi;s_0)\,|) \quad \text{as } \xi \to +\infty, \end{cases}$$

where M_{j0} $(j = 1,2,3,4)$ are some constants. By (3.7), we have $M_{20} = -C_u^+\,\Sigma_2^-(s_0) > 0$ and $M_{30} = -C_v^-\,\Lambda_1^+(s_0) < 0$.

We put $\boldsymbol{U}_4(\xi) = \Phi_2^+(\xi)$. Since $\Phi_2^+(\xi) = M_{44}\,\Psi_2^+(\xi) + o(|\,\Psi_2^+(\xi)\,|)$ as $\xi \to +\infty$, we see that the limits

$$\text{(3.9)} \qquad \lim_{\xi\to-\infty} \left|\, \boldsymbol{U}_4(\xi)\, e^{-\Lambda_2^+(s_0)\,\xi} \,\right|, \quad \lim_{\xi\to+\infty} \left|\, \boldsymbol{U}_4(\xi)\, \xi^{-n_+}\, e^{-\Sigma_2^+(s_0)\,\xi} \,\right|$$

exist and are nonzero.

We put

$$\boldsymbol{U}_2(\xi) = \Phi_2^-(\xi) - \frac{M_{42}}{M_{44}}\,\Phi_2^+(\xi), \quad \boldsymbol{U}_3(\xi) = \Phi_1^+(\xi) - \frac{M_{43}}{M_{44}}\,\Phi_2^+(\xi).$$

By $\Phi_1^+(\xi) = o(|\,\Phi_2^-(\xi)\,|)$ and $\Phi_2^+(\xi) = o(|\,\Phi_1^+(\xi)\,|)$ as $\xi \to -\infty$, we have

$$\boldsymbol{U}_2(\xi) = \Phi_2^-(\xi) + o(|\,\Phi_2^-(\xi)\,|), \quad \boldsymbol{U}_3(\xi) = \Phi_1^+(\xi) + o(|\,\Phi_1^+(\xi)\,|)$$

as $\xi \to -\infty$. Since $\boldsymbol{U}_2(\xi)$ and $\boldsymbol{U}_3(\xi)$ are represented as

$$\begin{aligned}
\boldsymbol{U}_2(\xi) &= N_{12}\,\Psi_1^-(\xi) + N_{22}\,\Psi_2^-(\xi) + N_{32}\,\Psi_1^+(\xi),\\
\boldsymbol{U}_3(\xi) &= N_{13}\,\Psi_1^-(\xi) + N_{23}\,\Psi_2^-(\xi) + N_{33}\,\Psi_1^+(\xi)
\end{aligned}$$

where

$$N_{ij} = \frac{M_{ij}\,M_{44} - M_{i4}\,M_{4j}}{M_{44}} \quad (i = 1,2,3;\, j = 2,3),$$

we also have $\boldsymbol{U}_2(\xi) = O(|\,\Psi_1^+(\xi)\,|)$ and $\boldsymbol{U}_3(\xi) = O(|\,\Psi_1^+(\xi)\,|)$ as $\xi \to +\infty$.

By $\Psi_1^+(\xi) = o(|\,\Psi_2^+(\xi)\,|)$ as $\xi \to +\infty$, we have

$$\begin{aligned}\boldsymbol{u}_{0\xi}(\xi) &= M_{30}\,\Phi_1^+(\xi) + M_{40}\,\Phi_2^+(\xi)\\ &= M_{30}\,\boldsymbol{U}_3(\xi) + \frac{M_{30}\,M_{43} + M_{40}\,M_{44}}{M_{44}}\,\Phi_2^+(\xi)\\ &= (M_{30}\,M_{43} + M_{40}\,M_{44})\,\Psi_2^+(\xi) + o(|\,\Psi_2^+(\xi)\,|)\end{aligned}$$

as $\xi \to +\infty$. From $\lim_{\xi\to+\infty} \boldsymbol{u}_{0\xi}(\xi) = 0$, we obtain $M_{30}\,M_{43} + M_{40}\,M_{44} = 0$, that is, $\boldsymbol{U}_3(\xi) = \boldsymbol{u}_{0\xi}(\xi)/M_{30}$. This means that the limits

$$\lim_{\xi\to-\infty}\left|\,\boldsymbol{U}_3(\xi)\,\xi^{-m_+}e^{-\Lambda_1^+(s_0)\,\xi}\,\right|,\quad \lim_{\xi\to+\infty}\left|\,\boldsymbol{U}_3(\xi)\,\xi^{-n_-}e^{-\Sigma_2^-(s_0)\,\xi}\,\right| \tag{3.10}$$

exist and are nonzero.

We assume that $N_{32} = 0$. By the use of $\boldsymbol{u}_{0\xi}(\xi) = M_{10}\,\Psi_1^-(\xi) + M_{20}\,\Psi_2^-(\xi)$ and $\Phi_2^-(\xi) = o(|\,\Phi_1^-(\xi)\,|)$ as $\xi \to -\infty$, we have

$$\begin{aligned}N_{22}\,\boldsymbol{u}_{0\xi}(\xi) - M_{20}\,\boldsymbol{U}_2(\xi) &= -\,M_{20}\,\Phi_2^-(\xi) + o(|\,\Phi_2^-(\xi)\,|) = o(|\,\Phi_1^-(\xi)\,|),\\ N_{22}\,\boldsymbol{u}_{0\xi}(\xi) - M_{20}\,\boldsymbol{U}_2(\xi) &= (M_{10}\,N_{22} - M_{20}\,N_{12})\,\Psi_1^-(\xi)\\ &= (M_{10}\,N_{22} - M_{20}\,N_{12})\,\widetilde{M}_{11}\\ &\quad\times(\Phi_1^-(\xi) + o(|\,\Phi_1^-(\xi)\,|))\end{aligned}$$

as $\xi \to -\infty$. From the above behavior, we have $M_{10}\,N_{22} - M_{20}\,N_{12} = 0$. By the asymptotic behavior of $\Phi_2^-(\xi)$, we obtain

$$0 = M_{20}\,[N_{22}\,\boldsymbol{u}_{0\xi}(\xi) - M_{20}\,\boldsymbol{U}_2(\xi)]_2 = -M_{20}^2\,e^{\Lambda_2^-(s_0)\,\xi}\,(1 + o(1)) < 0$$

as $\xi \to -\infty$. This contradiction implies that $N_{32} \neq 0$ holds. Then we see that the limits

$$\lim_{\xi\to-\infty}\left|\,\boldsymbol{U}_2(\xi)\,e^{-\Lambda_2^-(s_0)\,\xi}\,\right|,\quad \lim_{\xi\to+\infty}\left|\,\boldsymbol{U}_2(\xi)\,e^{-\Sigma_1^+(s_0)\,\xi}\,\right| \tag{3.11}$$

exist and are nonzero.

We put

$$\boldsymbol{U}_1(\xi) = \Phi_1^-(\xi) + \frac{\widetilde{M}_{21}}{\widetilde{M}_{11}}\,\Phi_2^-(\xi) + \frac{\widetilde{M}_{31}}{\widetilde{M}_{11}}\,\Phi_1^+(\xi) + \frac{\widetilde{M}_{41}}{\widetilde{M}_{11}}\,\Phi_2^+(\xi).$$

Then we have $\boldsymbol{U}_1(\xi) = \Psi_1^-(\xi)/\widetilde{M}_{11}$, that is, the limits

$$\lim_{\xi\to-\infty}\left|\,\boldsymbol{U}_1(\xi)\,\xi^{-m_-}\,e^{-\Lambda_1^-(s_0)\,\xi}\,\right|,\quad \lim_{\xi\to+\infty}\left|\,\boldsymbol{U}_1(\xi)\,e^{-\Sigma_1^-(s_0)\,\xi}\,\right| \tag{3.12}$$

exist and are nonzero.

By summarizing the above arguments, we have

$$\begin{aligned}&(\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3, \boldsymbol{U}_4)(\xi)\\ &\quad= (\Phi_1^-, \Phi_2^-, \Phi_1^+, \Phi_2^+)(\xi)\\ &\qquad\times\begin{pmatrix} 1 & 0 & 0 & 0\\ \widetilde{M}_{21}/\widetilde{M}_{11} & 1 & 0 & 0\\ \widetilde{M}_{31}/\widetilde{M}_{11} & 0 & 1 & 0\\ \widetilde{M}_{41}/\widetilde{M}_{11} & -M_{42}/M_{44} & -M_{43}/M_{44} & 1\end{pmatrix}.\end{aligned}$$

This formula shows that $\{\,\boldsymbol{U}_j(\xi)\,\}_{j=1}^4$ is a fundamental set of solutions of $\mathcal{L}(u, v) = 0$.

LEMMA 3.4. *There exists a fundamental set* $\{\boldsymbol{U}_j(\xi)\}_{j=1}^4$ *of solutions of* $\mathcal{L}(u,v) = 0$ *such that the limits* (3.9)–(3.12) *exist and are nonzero.*

3.3. Uniqueness. Let $\boldsymbol{u}_1(\xi) = (u_1, v_1)(\xi)$ and s_1 be an arbitrary positive solution and its propagation speed, respectively, of (1.5) for $(a,b,c) = (a_0, b_0, c_0)$. By the maximum principle and the positivity of $\boldsymbol{u}_1(\xi)$, we see that $\boldsymbol{u}_1(\xi)$ satisfies $\boldsymbol{u}_1(\xi) \in (0,1) \times (0, a_0)$ for any $\xi \in \mathbf{R}$ and has the same asymptotic expansion as $\boldsymbol{u}_0(\xi)$ near $\xi = \pm\infty$ with appropriate change of positive constants $C_u^{\pm}$ and $C_v^{\pm}$.

We first assume $s_0 < s_1$. Since (1.5) does not depend on ξ explicitly, we may assume $u_0(0) = 1/2 = u_1(0)$ without loss of generality. Since $\lambda_u^+(s)$, $\lambda_v^+(s)$, $\sigma_u^-(s)$, and $\sigma_v^-(s)$ are strictly decreasing with respect to s, we see that there exists $\xi_3 \geq 0$ such that $\boldsymbol{u}_1(\xi) \succ_o \boldsymbol{u}_0(\xi)$ for any $|\xi| \geq \xi_3$. We put

$$\underline{u}_1 = \min_{\xi\in[-\xi_3,\xi_3]} u_1(\xi)\,(\in (0,1)), \quad \underline{v}_1 = \max_{\xi\in[-\xi_3,\xi_3]} v_1(\xi)\,(\in (0,a_0)).$$

By $\boldsymbol{u}_0(-\infty) = (0, a_0)$, we can take $\xi_4\,(\in (-\infty, \xi_3])$ as satisfying $\boldsymbol{u}_0(\xi_4) \prec_o (\underline{u}_1, \underline{v}_1)$. Since $\boldsymbol{u}_0(\xi)$ is strictly monotone, we have

$$\boldsymbol{u}_1(\xi) - \boldsymbol{u}_0(\xi-\eta) \begin{cases} \succ_o \boldsymbol{u}_0(\xi) - \boldsymbol{u}_0(\xi-\eta) \succeq_o (0,0) & \text{if } |\xi| \geq \xi_3, \\ \succeq_o \boldsymbol{u}_1(\xi) - \boldsymbol{u}_0(\xi-(\xi_3-\xi_4)) & \\ \quad \succeq_o (\underline{u}_1, \underline{v}_1) - \boldsymbol{u}_0(\xi_4) \succ_o (0,0) & \text{if } |\xi| \leq \xi_3 \end{cases}$$

for any $\eta \geq \xi_3 - \xi_4\,(\geq 0)$. We put

$$\eta_1 = \sup\{\eta \mid \boldsymbol{u}_1(\xi) \not\succ_o \boldsymbol{u}_0(\xi-\eta) \text{ for some } \xi\}.$$

By virtue of $u_0(0) = 1/2 = u_1(0)$, we have $0 \leq \eta_1 \leq \xi_3 - \xi_4$ and

$$\boldsymbol{u}_2(\xi)(= (u_2, v_2)(\xi)) = \boldsymbol{u}_1(\xi) - \boldsymbol{u}_0(\xi-\eta_1) \succeq_o (0,0)$$

for any $\xi \in \mathbf{R}$. We assume that both $\eta_1 > 0$ and $\boldsymbol{u}_2(\xi) \succ_o (0,0)$ for any $\xi \in \mathbf{R}$ hold. Since $\boldsymbol{u}_1(\xi) - \boldsymbol{u}_0(\xi-\eta)$ is strictly monotone in η for any ξ, it follows that there exists $\eta_2 \in (0, \eta_1)$ such that $\boldsymbol{u}_1(\xi) - \boldsymbol{u}_0(\xi - (\eta_1 - \eta)) \succ_o (0,0)$ for any $|\xi| \leq \xi_3$ and $0 \leq \eta \leq \eta_2$. We obtain

$$\boldsymbol{u}_1(\xi) - \boldsymbol{u}_0(\xi - (\eta_1 - \eta)) \succ_o \boldsymbol{u}_0(\xi) - \boldsymbol{u}_0(\xi - (\eta_1 - \eta)) \succ_o (0,0)$$

for any $|\xi| \geq \xi_3$ and $0 \leq \eta \leq \eta_2$. This contradicts the definition of η_1. Since $u_2(0) = 0$ holds if $\eta_1 = 0$, we see that $u_2(\xi)$ and/or $v_2(\xi)$ attain local extremum 0 at some $\xi = \xi_5 \in \mathbf{R}$. Then we have

$$\begin{aligned} 0 \leq & u_{2\xi\xi}(\xi_5) \\ = & - s_1\, u_{1\xi}(\xi_5) - f(\boldsymbol{u}_1(\xi_5)) + s_0\, u_{0\xi}(\xi_5 - \eta_1) + f(\boldsymbol{u}_0(\xi_5 - \eta_1)) \\ = & (s_0 - s_1)\, u_{0\xi}(\xi_5 - \eta_1) + c_0\, u_0(\xi_5 - \eta_1)\, v_2(\xi_5) < 0 \end{aligned}$$

when $u_2(\xi_5) = 0$, and

$$0 \geq d\, v_{2\xi\xi}(\xi_5) = (s_0 - s_1)\, v_{0\xi}(\xi_5 - \eta_1) + b_0\, v_0(\xi_5 - \eta_1)\, u_2(\xi_5) > 0$$

when $v_2(\xi_5) = 0$. These contradictions imply that $s_0 \geq s_1$ holds.

Since we can also derive a contradiction in like manner when we assume $s_0 > s_1$, we have $s_0 = s_1$ as a result.

It follows from (3.7) that $\boldsymbol{u}_1(\xi)$ satisfies

$$\boldsymbol{u}_1(\xi) = (0, a_0) - C_2\, \overline{\Phi}_1^+(\xi; s_0) + o(|\,\overline{\Phi}_1^+(\xi; s_0)\,|)$$

as $\xi \to -\infty$, where C_2 is some positive constant. Since $\overline{\Phi}_1^+(\xi+\eta; s_0) = e^{\Lambda_1^+(s_0)\eta}\,\overline{\Phi}_1^+(\xi; s_0)$ $(1+o(1))$ as $\xi \to -\infty$ for any fixed η, we take η_3 as $C_2 = C_v^-\, e^{-\Lambda_1^+(s_0)\eta_3}$, and put

$$\boldsymbol{u}_3(\xi)(=(u_3, v_3)(\xi)) = \boldsymbol{u}_1(\xi) - \boldsymbol{u}_0(\xi - \eta_3).$$

By virtue of the choice of η_3, we have $\boldsymbol{u}_3(\xi) = o(|\,\overline{\Phi}_1^+(\xi; s_0)\,|)$ as $\xi \to -\infty$. And we see that $\boldsymbol{u}_3(\xi)$ satisfies

$$\begin{cases} u_{3\xi\xi} + s_0\, u_{3\xi} + f_u^1(\xi)\, u_3 + f_v^1(\xi)\, v_3 = 0, \\ d\, v_{3\xi\xi} + s_0\, v_{3\xi} + g_u^1(\xi)\, u_3 + g_v^1(\xi)\, v_3 = 0, \quad \xi \in \mathbf{R}, \end{cases} \tag{3.13}$$

where

$$f_u^1(\xi) = \int_0^1 f_u((1-\theta)\,\boldsymbol{u}_0(\xi-\eta_3) + \theta\,\boldsymbol{u}_1(\xi))\,d\theta$$

and the other functions f_v^1, g_u^1, and g_v^1 are also defined in the same manner with f_u^1. By Lemma 3.3 and

$$\begin{pmatrix} f_u^1(\xi) & f_v^1(\xi) \\ g_u^1(\xi) & g_v^1(\xi) \end{pmatrix} = \begin{pmatrix} 1 - a_0\, c_0 & 0 \\ -a_0\, b_0 & -a_0 \end{pmatrix} + o(e^{(\Lambda_1^+(s_0)-\delta)\,\xi})$$

as $\xi \to -\infty$ for any $0 < \delta < \Lambda_1^+(s_0)$, it follows that there exists a fundamental set $\{\,\widehat{\Phi}_1^{\pm}(\xi), \widehat{\Phi}_2^{\pm}(\xi)\,\}$ of solutions of (3.13) such that $\widehat{\Phi}_j^{\pm}(\xi) = \overline{\Phi}_j^{\pm}(\xi; s_0) + o(|\,\overline{\Phi}_j^{\pm}(\xi; s_0)\,|)$ as $\xi \to -\infty$. Since $\boldsymbol{u}_3(\xi)$ is a bounded solution of (3.13) and satisfies $\boldsymbol{u}_3(\xi) = o(|\,\overline{\Phi}_1^+(\xi; s_0)\,|)$ as $\xi \to -\infty$, we have $\boldsymbol{u}_3(\xi) = C_3\,\widehat{\Phi}_2^+(\xi)$ for some constant C_3. By

$$\begin{aligned} f_v^1(\xi) &= -\,c_0 \int_0^1 \{\,(1-\theta)\,u_0(\xi-\eta_3) + \theta\,u_1(\xi)\,\}\,d\theta \\ &= -\,\frac{c_0}{2}\,\{\,u_0(\xi-\eta_3) + u_1(\xi)\,\} = -c_0\,C_4\,e^{\lambda_u^+(s_0)\,\xi}\,(1+o(1)) \end{aligned}$$

as $\xi \to -\infty$, where C_4 is some positive constant, we obtain

$$\widehat{\Phi}_2^+(\xi)\,e^{-\Lambda_2^+(s_0)\,\xi} = \begin{cases} \left(\dfrac{p_v^-(\lambda_u^+(s_0); s_0)}{a_0\, b_0},\, 1\right)(1+o(1)) \\ \qquad\qquad \text{if } \lambda_u^+(s_0) > \lambda_v^+(s_0), \\ \left(\dfrac{c_0\, C_4\, e^{\lambda_u^+(s_0)\,\xi}}{p_u^-(\lambda_u^+(s_0) + \lambda_v^+(s_0); s_0)},\, 1\right)(1+o(1)) \\ \qquad\qquad \text{if } \lambda_u^+(s_0) \le \lambda_v^+(s_0) \end{cases}$$

as $\xi \to -\infty$ in a similar manner with the calculation of the asymptotic expansion of $\Phi_2^+(\xi)$. By virtue of (3.8), we have $\widehat{\Phi}_2^+(\xi) \succ_s (0,0)$ as $\xi \to -\infty$.

We assume $C_3 \neq 0$. Then we obtain $u_3(\xi)\,v_3(\xi) > 0$ near $\xi = -\infty$. We put

$$\xi_6 = \inf\{\,\xi \mid u_3(\xi)\,v_3(\xi) \le 0\,\}$$

if $u_3(\xi)$ and/or $v_3(\xi)$ have a zero, but otherwise $\xi_6 = +\infty$. Using (3.2) and the

integration by parts, we have

$$\begin{aligned}
0 &= \int_{-\infty}^{\xi_6} \{ u_{1\xi\xi}(\xi) + s_0\, u_{1\xi}(\xi) + f(\boldsymbol{u}_1(\xi)) \}\, u_0(\xi-\eta_3)\, e^{s_0\,\xi}\, d\xi \\
&= u_{1\xi}(\xi_6)\, u_0(\xi_6-\eta_3)\, e^{s_0\,\xi_6} - u_1(\xi_6)\, u_{0\xi}(\xi_6-\eta_3)\, e^{s_0\,\xi_6} \\
&\qquad + \int_{-\infty}^{\xi_6} \{ u_0(\xi-\eta_3)\, f(\boldsymbol{u}_1(\xi)) - u_1(\xi)\, f(\boldsymbol{u}_0(\xi-\eta_3)) \}\, e^{s_0\,\xi}\, d\xi \\
&= u_1(\xi_6)\, u_{3\xi}(\xi_6)\, e^{s_0\,\xi_6} \\
&\qquad - \int_{-\infty}^{\xi_6} u_0(\xi-\eta_3)\, u_1(\xi)\, \{ u_3(\xi) + c_0\, v_3(\xi) \}\, e^{s_0\,\xi}\, d\xi \neq 0
\end{aligned}$$

when $u_3(\xi_6) = 0$, and

$$\begin{aligned}
0 &= \int_{-\infty}^{\xi_6} \{ d\, v_{1\xi\xi}(\xi) + s_0\, v_{1\xi}(\xi) + g(\boldsymbol{u}_1(\xi)) \}\, v_0(\xi-\eta_3)\, e^{(s_0\,\xi)/d}\, d\xi \\
&= d\, v_1(\xi_6)\, v_{3\xi}(\xi_6)\, e^{(s_0\,\xi_6)/d} \\
&\qquad - \int_{-\infty}^{\xi_6} v_0(\xi-\eta_3)\, v_1(\xi)\, \{ b_0\, u_3(\xi) + v_3(\xi) \}\, e^{(s_0\,\xi)/d}\, d\xi \neq 0
\end{aligned}$$

when $v_3(\xi_6) = 0$ and/or $\xi_6 = +\infty$. The aforementioned contradictions imply that $\boldsymbol{u}_1(\xi) = \boldsymbol{u}_0(\xi-\eta_3)$ holds for any $\xi \in \mathbf{R}$.

LEMMA 3.5. *Let $\boldsymbol{u}_1(\xi)$ and s_1 be an arbitrary positive solution and its propagation speed, respectively, of* (1.5) *for $(a,b,c) = (a_0,b_0,c_0)$. Then $s_1 = s_0$ holds and there exists η such that $\boldsymbol{u}_1(\xi) = \boldsymbol{u}_0(\xi+\eta)$ for any $\xi \in \mathbf{R}$.*

3.4. Continuation.

LEMMA 3.6. *$\boldsymbol{u}_0(\xi)$ and s_0 satisfy*

$$\| \boldsymbol{u}_0 \|_{C^2(\mathbf{R})} \leq \begin{cases} C_5(b_0,c_0) & \text{for } s_0 = 0, \\ C_5(b_0,c_0)/|\, s_0\,| & \text{for } s_0 \neq 0, \end{cases}$$

and

$$-2 < s_0 < 2\sqrt{a_0\, d},$$

where $C_5(b_0,c_0)$ is a continuous function in (b_0,c_0) which does not depend on a_0.

Proof. We put

$$U(\xi) = \begin{cases} u_0(\xi)\, e^{(s_0\,\xi)/2} & \text{for } s_0 < 0, \\ v_0(\xi)\, e^{(s_0\,\xi)/(2\,d)} & \text{for } s_0 > 0. \end{cases}$$

We see from (3.2), (3.5), and (3.6) that $U(\xi)$ satisfies $\lim_{\xi\to\pm\infty} U(\xi) = 0$ for any $s_0 \neq 0$. Then it follows that there exists $\xi_7 \in \mathbf{R}$ such that $U(\xi)$ attains positive maximum at $\xi = \xi_7$, that is, $U(\xi)$ satisfies

$$0 \geq U_{\xi\xi}(\xi_7) = \left(\frac{s_0^2}{4} - 1 + u_0(\xi_7) + c_0\, v_0(\xi_7) \right) U(\xi_7)$$

when $s_0 < 0$, and

$$0 \geq U_{\xi\xi}(\xi_7) = \left(\frac{s_0^2}{4\,d} - a_0 + b_0\, u_0(\xi_7) + v_0(\xi_7) \right) \frac{U(\xi_7)}{d}$$

when $s_0 > 0$. Thus we have $-2 < s_0 < 2\sqrt{a_0\, d}$.

Since $0 < u_0(\xi) < 1$ and $0 < v_0(\xi) < a_0 < b_0$ hold for any $\xi \in \mathbf{R}$, we have $\| \boldsymbol{u}_0 \|_{C^0(\mathbf{R})} \leq 1 + b_0$ and

$$
\begin{aligned}
&-b_0\, c_0 < -c_0\, v_0(\xi) \leq f(\boldsymbol{u}_0(\xi)) \leq u_0(\xi) < 1, \\
&-b_0 < -b_0\, u_0(\xi) \leq g(\boldsymbol{u}_0(\xi)) \leq a_0\, v_0(\xi) < b_0^2
\end{aligned} \tag{3.14}
$$

for any $\xi \in \mathbf{R}$. We have the following:

(i) For $s_0 = 0$,

$$
\begin{aligned}
0 =& u_{0\xi}(\xi)^2 + 2 \int_{-\infty}^{\xi} f(\boldsymbol{u}_0(\tau))\, u_{0\xi}(\tau)\, d\tau \\
&\geq u_{0\xi}(\xi)^2 - 2\, b_0\, c_0 \int_{-\infty}^{\xi} u_{0\xi}(\tau)\, d\tau \geq u_{0\xi}(\xi)^2 - 2\, b_0\, c_0, \\
0 =& d\, v_{0\xi}(\xi)^2 + 2 \int_{-\infty}^{\xi} g(\boldsymbol{u}_0(\tau))\, v_{0\xi}(\tau)\, d\tau \geq d\, v_{0\xi}(\xi)^2 - 2\, b_0^3.
\end{aligned}
$$

(ii) For $s_0 > 0$,

$$
\begin{aligned}
0 =& u_{0\xi}(\xi) + \int_{-\infty}^{\xi} e^{s_0(\tau-\xi)} f(\boldsymbol{u}_0(\tau))\, d\tau \\
&\geq u_{0\xi}(\xi) - b_0\, c_0 \int_{-\infty}^{\xi} e^{s_0(\tau-\xi)}\, d\tau = u_{0\xi}(\xi) - b_0\, c_0/s_0, \\
0 =& d\, v_{0\xi}(\xi) + \int_{-\infty}^{\xi} e^{\{s_0(\tau-\xi)\}/d} g(\boldsymbol{u}_0(\tau))\, d\tau \leq d\, v_{0\xi}(\xi) + b_0^2\, d/s_0.
\end{aligned}
$$

(iii) For $s_0 < 0$,

$$
\begin{aligned}
0 =& u_{0\xi}(\xi) - \int_{\xi}^{+\infty} e^{s_0(\tau-\xi)} f(\boldsymbol{u}_0(\tau))\, d\tau, \\
&\geq u_{0\xi}(\xi) - \int_{\xi}^{+\infty} e^{s_0(\tau-\xi)}\, d\tau = u_{0\xi}(\xi) + 1/s_0, \\
0 =& d\, v_{0\xi}(\xi) - \int_{\xi}^{+\infty} e^{\{s_0(\tau-\xi)\}/d} g(\boldsymbol{u}_0(\tau))\, d\tau \leq d\, v_{0\xi}(\xi) - b_0\, d/s_0.
\end{aligned}
$$

By the above inequalities and $u_{0\xi}(\xi) > 0 > v_{0\xi}(\xi)$ for any $\xi \in \mathbf{R}$, we see that there exists a continuous function $C_6(b_0, c_0)$ independent of a_0 such that

$$
\| \boldsymbol{u}_{0\xi} \|_{C^0(\mathbf{R})} \leq \begin{cases} C_6(b_0, c_0) & \text{for } s_0 = 0, \\ C_6(b_0, c_0)/|\, s_0\,| & \text{for } s_0 \neq 0. \end{cases}
$$

By the inequalities $|\, s_0\,| < \max\{2, 2\sqrt{b_0\, d}\}$ and

$$
\| \boldsymbol{u}_{0\xi\xi} \|_{C^0(\mathbf{R})} \leq \frac{1}{\min(1, d)} \left\{ |\, s_0\,|\, \| \boldsymbol{u}_{0\xi} \|_{C^0(\mathbf{R})} + \| (f(\boldsymbol{u}_0), g(\boldsymbol{u}_0)) \|_{C^0(\mathbf{R})} \right\},
$$

we have the desired estimate for $\boldsymbol{u}_0(\xi)$. □

We put $(a, b, c, s) = (a_0, b_0, c_0, s_0) + (\widehat{a}, \widehat{b}, \widehat{c}, \widehat{s})$, $\boldsymbol{z} = {}^t(u, u_\xi, v, v_\xi)$ and

$$
\boldsymbol{f}(\boldsymbol{z}; \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s}) = \begin{pmatrix} u_\xi \\ -s\, u_\xi - u\,(1 - u - c\, v) \\ v_\xi \\ -\{\, s\, v_\xi + v\,(a - b\, u - v)\,\}/d \end{pmatrix}.
$$

Since (1.5) does not depend on ξ explicitly, we may assume $u(0) = 1/2$ without loss of generality. Then (1.5) is rewritten as

$$\text{(3.15)}\qquad \begin{cases} \dfrac{d}{d\xi}\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}; \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s}), \quad \xi \in \mathbf{R}, \\ \boldsymbol{z}(-\infty) = {}^t(0, 0, a_0 + \widehat{a}, 0), \\ \boldsymbol{z}(+\infty) = {}^t(1, 0, 0, 0), \\ \boldsymbol{z}(0) = \boldsymbol{z}_0(0) + {}^t(0, \alpha_1, \alpha_2, \alpha_3), \end{cases}$$

where $\boldsymbol{z}_0(\xi) = {}^t(u_0, u_{0\xi}, v_0, v_{0\xi})(\xi)$, and $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ will be determined later. Further, by the change of variables

$$\boldsymbol{z} = \begin{cases} \boldsymbol{z}_0(\xi) + {}^t(0, 0, \widehat{a}, 0) + \boldsymbol{y} & \text{for } \xi < 0, \\ \boldsymbol{z}_0(\xi) + \boldsymbol{y} & \text{for } \xi > 0, \end{cases}$$

(3.15) becomes

$$\text{(3.16)}\qquad \begin{cases} \dfrac{d}{d\xi}\boldsymbol{y} = A(\xi)\,\boldsymbol{y} + \boldsymbol{N}(\xi, \boldsymbol{y}; \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s}), \quad \xi \in \mathbf{R} \setminus \{0\}, \\ \boldsymbol{y}(-\infty) = 0 = \boldsymbol{y}(+\infty), \\ \boldsymbol{y}(0-0) = {}^t(0, \alpha_1, \alpha_2 - \widehat{a}, \alpha_3), \\ \boldsymbol{y}(0+0) = {}^t(0, \alpha_1, \alpha_2, \alpha_3), \end{cases}$$

where $A(\xi) = \boldsymbol{f}_{\boldsymbol{z}}(\boldsymbol{z}_0(\xi); 0)$,

$$\boldsymbol{N}(\xi, \boldsymbol{y}; \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s}) = \boldsymbol{f}(\boldsymbol{z}; \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s}) - \boldsymbol{f}(\boldsymbol{z}_0(\xi); 0) - A(\xi)\,\boldsymbol{y}.$$

Let J be an interval and $Y(\xi)$ be a fundamental matrix of $\frac{d}{d\xi}\boldsymbol{y} = A(\xi)\,\boldsymbol{y}$. We shall say that $\frac{d}{d\xi}\boldsymbol{y} = A(\xi)\,\boldsymbol{y}$ has an *exponential dichotomy* on J if there exist a projection P and positive constants C, γ such that, for any ξ, $\eta \in J$, $Y(\xi)$ satisfies

$$\begin{aligned} &|\,Y(\xi)\,P\,Y(\eta)^{-1}\,| \le C\,e^{-\gamma(\xi-\eta)} && \text{if } \xi \ge \eta, \\ &|\,Y(\xi)\,(I-P)\,Y(\eta)^{-1}\,| \le C\,e^{-\gamma(\eta-\xi)} && \text{if } \eta \ge \xi. \end{aligned}$$

We note that this definition does not depend on the choice of the fundamental matrix $Y(\xi)$.

PROPOSITION 3.7 (Coppel [3, p. 11]). *$\frac{d}{d\xi}\boldsymbol{y} = A(\xi)\,\boldsymbol{y}$ has an exponential dichotomy on J if and only if there exist a projection P and positive constants C, γ such that, for any ξ, $\eta \in J$, $Y(\xi)$ satisfies*

$$\begin{aligned} &|\,Y(\xi)\,P\,\boldsymbol{w}\,| \le C\,e^{-\gamma(\xi-\eta)}\,|\,Y(\eta)\,P\,\boldsymbol{w}\,| && \text{if } \xi \ge \eta, \\ &|\,Y(\xi)\,(I-P)\,\boldsymbol{w}\,| \le C\,e^{-\gamma(\eta-\xi)}\,|\,Y(\eta)\,(I-P)\,\boldsymbol{w}\,| && \text{if } \eta \ge \xi, \end{aligned}$$

where $\boldsymbol{w}$ is an arbitrary constant vector.

We define the matrix $X(\xi)$ by

$$X(\xi) = \begin{pmatrix} U_3(\xi) & U_4(\xi) & U_1(\xi) & U_2(\xi) \\ U_{3\xi}(\xi) & U_{4\xi}(\xi) & U_{1\xi}(\xi) & U_{2\xi}(\xi) \\ V_3(\xi) & V_4(\xi) & V_1(\xi) & V_2(\xi) \\ V_{3\xi}(\xi) & V_{4\xi}(\xi) & V_{1\xi}(\xi) & V_{2\xi}(\xi) \end{pmatrix},$$

where $\{\,(U_j, V_j)(\xi)\,\}_{j=1}^4$ is a fundamental set of solutions of $\mathcal{L}(u, v) = 0$ given in Lemma 3.4. In consideration of the proof of Lemma 3.4, we can take $(U_3, V_3)(\xi)$ as

$(U_3, V_3)(\xi) = \boldsymbol{u}_{0\xi}(\xi)$. Clearly we see that $X(\xi)$ is a fundamental matrix of $\frac{d}{d\xi}\boldsymbol{y} = A(\xi)\boldsymbol{y}$. By virtue of (3.9)–(3.12), the following lemma holds.

LEMMA 3.8. *$\frac{d}{d\xi}\boldsymbol{y} = A(\xi)\boldsymbol{y}$ has an exponential dichotomy on $R_- = (-\infty, 0)$ (respectively, $R_+ = (0, +\infty)$) with the projection $P_- = \mathrm{diag}(0,0,1,1)$ (respectively, $P_+ = \mathrm{diag}(1,0,1,0)$).*

Proof. We may take the constant γ in Proposition 3.7 as satisfying $0 < \gamma < \min\{\Lambda_1^+(s_0), -\Sigma_2^-(s_0)\}$. □

Let $\boldsymbol{x}_j^*(\xi)$ and $x_{ij}^*(\xi)$ be jth row vector and (i,j) element, respectively, of $X(\xi)^{-1}$. We calculate $X(\xi)^{-1}$ directly by the use of the asymptotic behaviors (3.9)–(3.12), and then obtain

$$\begin{aligned}
\boldsymbol{x}_2^*(\xi) &= o(e^{-(\Sigma_2^+(s_0)-\delta)\xi}) \quad \text{as } \xi \to +\infty,\\
\boldsymbol{x}_3^*(\xi) &= o(e^{-(\Lambda_1^-(s_0)+\delta)\xi}) \quad \text{as } \xi \to -\infty,\\
\boldsymbol{x}_4^*(\xi) &= \begin{cases} o(e^{-(\Lambda_2^-(s_0)+\delta)\xi}) & \text{as } \xi \to -\infty,\\ o(e^{-(\Sigma_1^+(s_0)-\delta)\xi}) & \text{as } \xi \to +\infty \end{cases}
\end{aligned}$$

for any $0 < \delta < \min\{-\Lambda_2^-(s_0), \Sigma_1^+(s_0)\}$. Then we see that $(u^*, v^*)(\xi) = (x_{42}^*, x_{44}^*/d)(\xi)$ is a bounded nontrivial solution of $\mathcal{L}^*(u,v) = 0$.

LEMMA 3.9 (Lemma A.2 in [8]). *$(u^*, v^*)(\xi)$ satisfies $u^*(\xi)\, v^*(\xi) < 0$ for any $\xi \in \mathbf{R}$.*

LEMMA 3.10 (Lemma 3.2 in Kokubu [9]). *$\boldsymbol{y}(\xi)$ is a bounded solution of*

$$\begin{cases} \dfrac{d}{d\xi}\boldsymbol{y} = A(\xi)\,\boldsymbol{y} + \boldsymbol{N}(\xi, \boldsymbol{y}; \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s}), \quad \xi \in \mathbf{R} \setminus \{0\}, \\ \boldsymbol{y}(-\infty) = 0, \quad \boldsymbol{y}(+\infty) = 0 \end{cases}$$

if and only if $\boldsymbol{y}(\xi)$ satisfies

$$P_- \left\{ X(0)^{-1}\boldsymbol{y}(0-0) - \int_{R_-} X(\xi)^{-1}\boldsymbol{N}(\xi, \boldsymbol{y}(\xi); \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s})\, d\xi \right\} = 0$$

and

$$(I - P_+) \left\{ X(0)^{-1}\boldsymbol{y}(0+0) + \int_{R_+} X(\xi)^{-1}\boldsymbol{N}(\xi, \boldsymbol{y}(\xi); \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s})\, d\xi \right\} = 0.$$

From the previous lemma and (3.16), we define the map $E : \mathbf{R}^7 \to \mathbf{R}^4$ by

$$E(\alpha, \widehat{s}, \widehat{a}, \widehat{b}, \widehat{c}) = \begin{pmatrix} \boldsymbol{x}_3^*(0)\,\boldsymbol{y}(0-0) - \displaystyle\int_{R_-} \boldsymbol{x}_3^*(\xi)\,\boldsymbol{N}(\xi, \boldsymbol{y}(\xi); \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s})\, d\xi \\ \boldsymbol{x}_4^*(0)\,\boldsymbol{y}(0-0) - \displaystyle\int_{R_-} \boldsymbol{x}_4^*(\xi)\,\boldsymbol{N}(\xi, \boldsymbol{y}(\xi); \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s})\, d\xi \\ \boldsymbol{x}_2^*(0)\,\boldsymbol{y}(0+0) + \displaystyle\int_{R_+} \boldsymbol{x}_2^*(\xi)\,\boldsymbol{N}(\xi, \boldsymbol{y}(\xi); \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s})\, d\xi \\ \boldsymbol{x}_4^*(0)\,\boldsymbol{y}(0+0) + \displaystyle\int_{R_+} \boldsymbol{x}_4^*(\xi)\,\boldsymbol{N}(\xi, \boldsymbol{y}(\xi); \widehat{a}, \widehat{b}, \widehat{c}, \widehat{s})\, d\xi \end{pmatrix}.$$

We easily obtain $E(0) = 0$. Since $\boldsymbol{x}_j^*(\xi)$ is a solution of $\frac{d}{d\xi}\boldsymbol{x} = -\boldsymbol{x}\,A(\xi)$ for each j, we

have

$$\frac{\partial}{\partial\mu}\left[\boldsymbol{x}_j^*(0)\,\boldsymbol{y}(0-0)-\int_{R_-}\boldsymbol{x}_j^*(\xi)\,\boldsymbol{N}(\xi,\boldsymbol{y}(\xi);\widehat{a},\widehat{b},\widehat{c},\widehat{s})\,d\xi\right]\Bigg|_{(\alpha,\widehat{a},\widehat{b},\widehat{c},\widehat{s})=0}$$
$$=-\delta_{\mu,\widehat{a}}\,\boldsymbol{x}_j^*(0)\,\boldsymbol{e}-\int_{R_-}\boldsymbol{x}_j^*(\xi)\,\{\,\delta_{\mu,\widehat{a}}\,\boldsymbol{f}_{\boldsymbol{z}}(\boldsymbol{z}_0(\xi);0)\,\boldsymbol{e}+\boldsymbol{f}_\mu(\boldsymbol{z}_0(\xi);0)\,\}d\xi$$
$$=-\int_{R_-}\boldsymbol{x}_j^*(\xi)\,\boldsymbol{f}_\mu(\boldsymbol{z}_0(\xi);0)\,d\xi\quad\text{for } j=3,4,$$
$$\frac{\partial}{\partial\mu}\left[\boldsymbol{x}_j^*(0)\,\boldsymbol{y}(0+0)+\int_{R_+}\boldsymbol{x}_j^*(\xi)\,\boldsymbol{N}(\xi,\boldsymbol{y}(\xi);\widehat{a},\widehat{b},\widehat{c},\widehat{s})\,d\xi\right]\Bigg|_{(\alpha,\widehat{a},\widehat{b},\widehat{c},\widehat{s})=0}$$
$$=\int_{R_+}\boldsymbol{x}_j^*(\xi)\,\boldsymbol{f}_\mu(\boldsymbol{z}_0(\xi);0)\,d\xi\quad\text{for } j=2,4$$

with $\mu=\widehat{a},\widehat{b},\widehat{c},\widehat{s}$, where $\boldsymbol{e}={}^t(0,0,1,0)$, and $\delta_{\mu,\widehat{a}}$ satisfies $\delta_{\mu,\widehat{a}}=1$ if $\mu=\widehat{a}$ but otherwise $\delta_{\mu,\widehat{a}}=0$. By the above formulas, we obtain

$$\det\frac{\partial E}{\partial(\alpha,\widehat{s})}(0)$$
$$=\det\begin{pmatrix} x_{32}^*(0) & x_{33}^*(0) & x_{34}^*(0) & -\displaystyle\int_{R_+}\boldsymbol{x}_3^*(\xi)\,\boldsymbol{f}_{\widehat{s}}(\boldsymbol{z}_0(\xi);0)\,d\xi\\ x_{42}^*(0) & x_{43}^*(0) & x_{44}^*(0) & -\displaystyle\int_{R_-}\boldsymbol{x}_4^*(\xi)\,\boldsymbol{f}_{\widehat{s}}(\boldsymbol{z}_0(\xi);0)\,d\xi\\ x_{22}^*(0) & x_{23}^*(0) & x_{24}^*(0) & \displaystyle\int_{R_+}\boldsymbol{x}_2^*(\xi)\,\boldsymbol{f}_{\widehat{s}}(\boldsymbol{z}_0(\xi);0)\,d\xi\\ x_{42}^*(0) & x_{43}^*(0) & x_{44}^*(0) & \displaystyle\int_{R_+}\boldsymbol{x}_4^*(\xi)\,\boldsymbol{f}_{\widehat{s}}(\boldsymbol{z}_0(\xi);0)\,d\xi \end{pmatrix}$$
$$=\frac{u_{0\xi}(0)}{\det X(0)}\int_{\mathbf{R}}\boldsymbol{x}_4^*(\xi)\,\boldsymbol{f}_{\widehat{s}}(\boldsymbol{z}_0(\xi);0)\,d\xi$$
$$=-\frac{u_{0\xi}(0)}{\det X(0)}\int_{\mathbf{R}}\{u_{0\xi}(\xi)\,u^*(\xi)+v_{0\xi}(\xi)\,v^*(\xi)\}\,d\xi.$$

We also get analogously

$$\det\frac{\partial E}{\partial(\alpha,\widehat{a})}(0)=-\frac{u_{0\xi}(0)}{\det X(0)}\int_{\mathbf{R}}v_0(\xi)\,v^*(\xi)\,d\xi,$$
$$\det\frac{\partial E}{\partial(\alpha,\widehat{b})}(0)=\frac{u_{0\xi}(0)}{\det X(0)}\int_{\mathbf{R}}u_0(\xi)\,v_0(\xi)\,v^*(\xi)\,d\xi,$$
$$\det\frac{\partial E}{\partial(\alpha,\widehat{c})}(0)=\frac{u_{0\xi}(0)}{\det X(0)}\int_{\mathbf{R}}u_0(\xi)\,v_0(\xi)\,u^*(\xi)\,d\xi.$$

We have the following lemmas by virtue of the implicit function theorem, the maximum principle, and Lemma 3.9.

LEMMA 3.11. *There exist C^1-class families $\overline{\boldsymbol{u}}(\xi;b,c)$ and $\overline{a}(b,c)$ defined in a neighborhood of $(b,c)=(b_0,c_0)$ such that $\overline{\boldsymbol{u}}(\xi;b,c)$ is a strictly monotone solution of (1.5) with $(s,a)=(s_0,\overline{a}(b,c))$ for each (b,c) and*

$$(\overline{\boldsymbol{u}}(.;b,c),\overline{a}(b,c))\to(\boldsymbol{u}_0(.),a_0)\quad as\quad(b,c)\to(b_0,c_0)$$

holds.

LEMMA 3.12. *There exist C^1-class families $\boldsymbol{u}(\xi;a,b,c)$ and $s(a,b,c)$ defined in a neighborhood of $(a,b,c)=(a_0,b_0,c_0)$ such that $\boldsymbol{u}(\xi;a,b,c)$ is a strictly monotone solution of* (1.5) *with $s=s(a,b,c)$ for each (a,b,c) and*

$$(\boldsymbol{u}(.;a,b,c),s(a,b,c))\to(\boldsymbol{u}_0(.),s_0)\quad as\quad (a,b,c)\to(a_0,b_0,c_0)$$

holds. Furthermore, $s(a,b,c)$ satisfies

$$\frac{\partial}{\partial a}s(a_0,b_0,c_0)=-\frac{\int_{\mathbf{R}}v_0(\xi)\,v^*(\xi)\,d\xi}{\int_{\mathbf{R}}\{u_{0\xi}(\xi)\,u^*(\xi)+v_{0\xi}(\xi)\,v^*(\xi)\}\,d\xi}>0,$$

$$\frac{\partial}{\partial b}s(a_0,b_0,c_0)=\frac{\int_{\mathbf{R}}u_0(\xi)\,v_0(\xi)\,v^*(\xi)\,d\xi}{\int_{\mathbf{R}}\{u_{0\xi}(\xi)\,u^*(\xi)+v_{0\xi}(\xi)\,v^*(\xi)\}\,d\xi}<0,$$

$$\frac{\partial}{\partial c}s(a_0,b_0,c_0)=\frac{\int_{\mathbf{R}}u_0(\xi)\,v_0(\xi)\,u^*(\xi)\,d\xi}{\int_{\mathbf{R}}\{u_{0\xi}(\xi)\,u^*(\xi)+v_{0\xi}(\xi)\,v^*(\xi)\}\,d\xi}>0.$$

3.5. Nonexistence. Let $\boldsymbol{u}(\xi)=(u,v)(\xi)$ be an arbitrary monotone bounded solution of (1.5a). Since $\boldsymbol{u}(\xi)$ is bounded and satisfies $u_\xi(\xi)\geq 0\geq v_\xi(\xi)$ for any $\xi\in\mathbf{R}$, we have $\liminf_{\xi\to\pm\infty}u_\xi(\xi)=0$ and $\limsup_{\xi\to\pm\infty}v_\xi(\xi)=0$. By (1.5a), we have

$$u_\xi(\tau)=u_\xi(\xi)+s\{u(\xi)-u(\tau)\}+\int_\tau^\xi f(\boldsymbol{u}(\eta))\,d\eta. \tag{3.17}$$

for any $\xi,\tau\in\mathbf{R}$. If $f(\boldsymbol{u}(\xi))\geq 0$ near $\xi=-\infty$ (respectively, $\xi=+\infty$), then we see that the limit $\int_{-\infty}^{\xi}f(\boldsymbol{u}(\eta))\,d\eta$ (respectively, $\int_{+\infty}^{\xi}f(\boldsymbol{u}(\eta))\,d\eta$) exists on $\overline{\mathbf{R}}$ for any ξ, because $\int_\tau^\xi f(\boldsymbol{u}(\eta))\,d\eta$ is increasing (respectively, decreasing) in τ for any ξ. By taking the inferior limit of both hands of (3.17) as $\tau\to-\infty$ (respectively, $\tau\to+\infty$), we get

$$\begin{aligned}&0=u_\xi(\xi)+s\{u(\xi)-u(-\infty)\}+\int_{-\infty}^{\xi}f(\boldsymbol{u}(\eta))\,d\eta\\&\left(\text{respectively, }0=u_\xi(\xi)+s\{u(\xi)-u(+\infty)\}-\int_{\xi}^{+\infty}f(\boldsymbol{u}(\eta))\,d\eta\right)\end{aligned} \tag{3.18}$$

for any $\xi\in\mathbf{R}$ if $f(\boldsymbol{u}(\xi))\geq 0$ near $\xi=-\infty$ (respectively, $\xi=+\infty$). We also have, analogously,

$$\begin{aligned}&0=dv_\xi(\xi)+s\{v(\xi)-v(-\infty)\}+\int_{-\infty}^{\xi}g(\boldsymbol{u}(\eta))\,d\eta\\&\left(\text{respectively, }0=dv_\xi(\xi)+s\{v(\xi)-v(+\infty)\}-\int_{\xi}^{+\infty}g(\boldsymbol{u}(\eta))\,d\eta\right)\end{aligned} \tag{3.19}$$

for any $\xi\in\mathbf{R}$ if $g(\boldsymbol{u}(\xi))\leq 0$ near $\xi=-\infty$ (respectively, $\xi=+\infty$).

LEMMA 3.13. *For each $(a,b,c)\in R^3_+$, if $s\geq 0$ (respectively, $s\leq 0$), then an arbitrary, monotone nonnegative solution $(u,v)(\xi)$ of* (1.5a) *with $(u,v)(-\infty)=(0,0)$ (respectively, $(u,v)(+\infty)=(0,0)$) satisfies $u(\xi)=0$ (respectively, $v(\xi)=0$) for any $\xi\in\mathbf{R}$.*

Proof. We only show the proof for the former case, because the latter can be proved in a similar manner.

Let $\boldsymbol{u}(\xi)=(u,v)(\xi)$ be an arbitrary monotone nonnegative solution of (1.5a) with $(u,v)(-\infty)=(0,0)$ for $s\geq 0$. We assume that $u(\xi_8)>0$ holds for some $\xi_8\in\mathbf{R}$. Since

$f(u,v) > 0$ holds on $(0,\delta] \times [0,\delta]$ for some $\delta > 0$, we see from $\boldsymbol{u}(-\infty) = (0,0)$ that there exists $\xi_9 \in \mathbf{R}$ such that $f(\boldsymbol{u}(\xi)) \geq 0\,(\not\equiv 0)$ for any $\xi \leq \xi_9$. By (3.18), we have

$$0 = u_\xi(\xi_9) + s\,u(\xi_9) + \int_{-\infty}^{\xi_9} f(\boldsymbol{u}(\xi))\,d\xi > 0.$$

This contradiction implies that the desired result holds. □

LEMMA 3.14. *Equation* (1.5) *with* $s = 0$ *has no monotone solution for arbitrary* $(a,b,c) \notin \mathcal{P}$.

Proof. We assume that (1.5) with $s = 0$ has a monotone solution $\boldsymbol{u}(\xi) = (u,v)(\xi)$ for some (a,b,c), which satisfies either $a < 1/c$ or $a = 1/c \leq b$.

We first consider the case for $a = 1/c \leq b$. Then we have $(1 - u - c\,v)(\pm\infty) = 0$. We assume that $(1 - u - c\,v)(\xi)$ attains a negative local minimum at $\xi = \xi_{10}$. By $f(\boldsymbol{u}(\xi_{10})) < 0$ and $g(\boldsymbol{u}(\xi_{10})) < 0$, we have

$$0 \leq (1 - u - c\,v)_{\xi\xi}(\xi_{10}) = f(\boldsymbol{u}(\xi_{10})) + \frac{c}{d}\,g(\boldsymbol{u}(\xi_{10})) < 0.$$

This contradiction implies that $f(\boldsymbol{u}(\xi)) \geq 0$ holds for any $\xi \in \mathbf{R}$. Then, by $\liminf_{\xi\to\pm\infty} u_\xi(\xi) = 0$ and $u_{\xi\xi}(\xi) = -f(\boldsymbol{u}(\xi)) \leq 0$ for any $\xi \in \mathbf{R}$, we obtain $u_\xi(\xi) = 0$ for any $\xi \in \mathbf{R}$, that is, $u(\xi)$ is a constant function. This contradicts the boundary conditions at $\xi = \pm\infty$.

Next, we consider the case for $a < 1/c$. Since $(1 - u - c\,v)(\xi) = 1 - a\,c + o(1) > 0$ as $\xi \to -\infty$, we have

$$0 = u_\xi(\xi) + \int_{-\infty}^{\xi} f(\boldsymbol{u}(\xi))\,d\xi > 0$$

as $\xi \to -\infty$ because of (3.18). This is a contradiction.

Thus we see that the desired result holds when (a,b,c) satisfies either $a < 1/c$ or $a = 1/c \leq b$.

In a similar manner, we can also prove the desired result for the case where either $a > b$ or $a = b \geq 1/c$ holds. □

LEMMA 3.15. *For arbitrary* $(a,b,c) \in \mathcal{P}$, *if* $s \geq 0$ (*respectively,* $s \leq 0$), *then* (1.5a) *has no monotone solution which satisfies*

$$(u,v)(-\infty) = \left(\frac{1 - a\,c}{1 - b\,c}, \frac{a - b}{1 - b\,c}\right), \quad (u,v)(+\infty) = (1,0)$$

$$\left(\textit{respectively,} \quad (u,v)(-\infty) = (0,a), \quad (u,v)(+\infty) = \left(\frac{1 - a\,c}{1 - b\,c}, \frac{a - b}{1 - b\,c}\right)\right).$$

Proof. We only show the proof for the former case, because the latter can be proved in a similar manner.

Contrary to the conclusion, suppose that (1.5a) with the above boundary conditions has a monotone solution $\boldsymbol{u}(\xi) = (u,v)(\xi)$ for some $s \geq 0$. By the boundary conditions at $\xi = \pm\infty$, we have

$$(1 - u - c\,v)(\pm\infty) = 0, \quad (a - b\,u - v)(\pm\infty) \leq 0.$$

We consider the case for $d \leq 1$, and assume that $(1 - u - c\,v)(\xi)$ attains a nonpositive local minimum at $\xi = \xi_{11}$. Since $u_\xi(\xi_{11}) + c\,v_\xi(\xi_{11}) = 0$ and $g(\boldsymbol{u}(\xi_{11})) < 0$,

we have

$$\begin{aligned}0 \leq & (1-u-cv)_{\xi\xi}(\xi_{11}) \\ = & s\left(1-\frac{1}{d}\right)u_\xi(\xi_{11}) + f(\boldsymbol{u}(\xi_{11})) + \frac{c}{d}\,g(\boldsymbol{u}(\xi_{11})) < 0.\end{aligned}$$

This contradiction implies that $f(\boldsymbol{u}(\xi)) > 0$ holds for any $\xi \in \mathbf{R}$ when $d \leq 1$. Analogously, we obtain $g(\boldsymbol{u}(\xi)) < 0$ for any $\xi \in \mathbf{R}$ when $d \geq 1$.

By (3.18) and (3.19), we have

$$0 = s\{u(+\infty) - u(-\infty)\} + \int_{\mathbf{R}} f(\boldsymbol{u}(\xi))\,d\xi > 0 \quad \text{for } d \leq 1,$$

$$0 = s\{v(+\infty) - v(-\infty)\} + \int_{\mathbf{R}} g(\boldsymbol{u}(\xi))\,d\xi < 0 \quad \text{for } d \geq 1.$$

These contradictions imply that the desired result holds. □

3.6. Proof of Theorem 2.1.

LEMMA 3.16. *Let* $\boldsymbol{u}(\xi) = (u,v)(\xi)$ *be a monotone, nonnegative, and bounded solution of* (1.5a) *for* $(a,b,c,s) \in R^3_+ \times \mathbf{R}$. *Then* $\boldsymbol{u}(\xi)$ *satisfies* $f(\boldsymbol{u}(\pm\infty)) = 0$ *and* $g(\boldsymbol{u}(\pm\infty)) = 0$.

Proof. We only show the proof for $f(\boldsymbol{u}(+\infty)) = 0$, because $f(\boldsymbol{u}(-\infty)) = 0$ and $g(\boldsymbol{u}(\pm\infty)) = 0$ can be proved in a similar manner.

Because $\boldsymbol{u}(\xi)$ is monotone bounded, we have $u_\xi(\xi) \geq 0$ for any $\xi \in \mathbf{R}$ and $\liminf_{\xi\to+\infty} u_\xi(\xi) = 0$. We assume $|f(\boldsymbol{u}(+\infty))|\,(\equiv 2f_+) > 0$. Then it follows that there exists ξ_{12} such that

$$f(\boldsymbol{u}(\xi))\begin{cases} \geq f_+ & \text{if } f(\boldsymbol{u}(+\infty)) > 0, \\ \leq -f_+ & \text{if } f(\boldsymbol{u}(+\infty)) < 0 \end{cases}$$

for any $\xi \geq \xi_{12}$.

We first consider the case where both $s \geq 0$ and $f(\boldsymbol{u}(+\infty)) > 0$ hold. Then we have

$$u_{\xi\xi}(\xi) = -s\,u_\xi(\xi) - f(\boldsymbol{u}(\xi)) \leq -f(\boldsymbol{u}(\xi)) \leq -f_+$$

for any $\xi \geq \xi_{12}$, that is,

$$u_\xi(\xi) \leq u_\xi(\xi_{12}) - (\xi - \xi_{12})\,f_+ \to -\infty$$

as $\xi \to +\infty$. This is a contradiction.

We next consider the case where both $s \geq 0$ and $f(\boldsymbol{u}(+\infty)) < 0$ hold. Then we have

$$u_{\xi\xi}(\xi) + s\,u_\xi(\xi) = -f(\boldsymbol{u}(\xi)) \geq f_+$$

for any $\xi \geq \xi_{12}$, that is, $u_\xi(\xi)$ satisfies

$$u_\xi(\xi) \geq e^{s(\xi_{12}-\xi)}\,u_\xi(\xi_{12}) + f_+ \int_{\xi_{12}}^{\xi} e^{s(\eta-\xi)}\,d\eta \to \begin{cases} +\infty & \text{if } s = 0, \\ f_+/s & \text{if } s > 0 \end{cases}$$

as $\xi \to +\infty$. This contradicts the fact that $\liminf_{\xi\to+\infty} u_\xi(\xi) = 0$.

We finally consider the case for $s < 0$. By (1.5a), we have

$$u_\xi(\xi) = e^{s(\tau-\xi)}\,u_\xi(\tau) + \int_{\xi}^{\tau} e^{s(\eta-\xi)}\,f(\boldsymbol{u}(\eta))\,d\eta$$

for any $\xi, \tau \in \mathbf{R}$. By taking the inferior limit of both hands as $\tau \to +\infty$, we obtain

$$u_\xi(\xi) = \int_\xi^{+\infty} e^{s(\eta-\xi)} f(\boldsymbol{u}(\eta))\, d\eta \begin{cases} \geq -f_+/s > 0 & \text{if } f(\boldsymbol{u}(+\infty)) > 0, \\ < 0 & \text{if } f(\boldsymbol{u}(+\infty)) < 0 \end{cases}$$

for any $\xi \geq \xi_{12}$. This is a contradiction.

Thus we have $f(\boldsymbol{u}(+\infty)) = 0$. □

Proof of Theorem 2.1. Let $\mathcal{P}_0 = \{\, (b,c) \,|\, 0 < 1/c < b \,\}$, and let β_0 be a constant which satisfies (3.1). We denote by $\mathcal{E}_0(\beta_0)\,(\subset R^2_+)$ the maximal extended and connected region which is given by applying Lemmas 3.1 and 3.11 to $(b,c) = (\beta_0, \beta_0)$ and $s_0 = 0$, and on which the functions $\overline{\boldsymbol{u}}(\xi; b, c, \beta_0) = (\overline{u}, \overline{v})(\xi; b, c, \beta_0)$ and $\overline{a}(b, c, \beta_0)$ given in Lemma 3.11 for $s_0 = 0$ are C^1 class. We see from $(\beta_0, \beta_0) \in \mathcal{E}_0(\beta_0)$ that $\mathcal{E}_0(\beta_0)$ is a nonempty set. By Lemma 3.14, we have $\mathcal{E}_0(\beta_0) \subset \mathcal{P}_0$ and $1/c < \overline{a}(b, c, \beta_0) < b$ for any $(b,c) \in \mathcal{E}_0(\beta_0)$.

We assume $\partial\mathcal{E}_0(\beta_0) \cap \mathcal{P}_0 \neq \emptyset$. Let $(b_0, c_0) \in \partial\mathcal{E}_0(\beta_0) \cap \mathcal{P}_0$, and let $\{\, (b_n, c_n) \,\}_{n=1}^{\infty} (\subset \mathcal{E}_0(\beta_0))$ be an arbitrary sequence which satisfies $(b_n, c_n) \to (b_0, c_0)$ as $n \to \infty$. Since $1/c_n < \overline{a}(b_n, c_n, \beta_0) < b_n$ for any integer n, we may assume $\overline{a}(b_n, c_n, \beta_0) \to \overline{a}_0 \in [1/c_0, b_0]$ as $n \to \infty$. And we may also assume

$$\overline{u}(0; b_n, c_n, \beta_0) = \begin{cases} \dfrac{2 - \{\overline{a}(b_n, c_n) + b_n\}\, c_n}{2\,(1 - b_n\, c_n)} & \text{if } \overline{a}_0 \in (1/c_0, b_0), \\ 1/2 & \text{if } \overline{a}_0 \in \{\, 1/c_0, b_0 \,\} \end{cases} \tag{3.20}$$

without loss of generality. Because (1.5) does not depend on ξ explicitly. It follows from the Ascoli–Arzela theorem and Lemma 3.6 that, for any fixed $\xi_{13} > 0$, there exist $\{\, (b_{n_j}, c_{n_j}) \,\}_{j=1}^{\infty} (\subset \{\, (b_n, c_n) \,\}_{n=1}^{\infty})$ and $\overline{\boldsymbol{u}}_0(\xi)$ such that

$$\lim_{j \to \infty} \| \overline{\boldsymbol{u}}(.; b_{n_j}, c_{n_j}, \beta_0) - \overline{\boldsymbol{u}}_0 \|_{C^2([-\xi_{13}, \xi_{13}])} = 0.$$

Since $\overline{\boldsymbol{u}}(\xi; b_{n_j}, c_{n_j}, \beta_0)$ is strictly monotone and satisfies

$$\overline{\boldsymbol{u}}(\xi; b_{n_j}, c_{n_j}, \beta_0) \in (0,1) \times (0, \overline{a}(b_{n_j}, c_{n_j}, \beta_0)) \quad \text{on } \mathbf{R}$$

for each j, we see that $\overline{\boldsymbol{u}}_0(\xi)$ is a monotone solution of (1.5a) with $s = 0$ for $(a,b,c) = (\overline{a}_0, b_0, c_0)$, which satisfies $\overline{\boldsymbol{u}}_0(\xi) \in [0,1] \times [0, \overline{a}_0]$ for any $\xi \in \mathbf{R}$.

We first consider the case where either $\overline{a}_0 = 1/c_0$ or $\overline{a}_0 = b_0$ holds. Then we see that the equilibrium points of (1.5a) are $(0,0)$, $(0, \overline{a}_0)$, and $(1,0)$. By $\overline{u}_0(0) = 1/2$ and Lemma 3.16, we have $\overline{\boldsymbol{u}}(+\infty) = (1,0)$ and either $\overline{\boldsymbol{u}}(-\infty) = (0,0)$ or $\overline{\boldsymbol{u}}(-\infty) = (0, \overline{a}_0)$. This contradicts the fact of Lemmas 3.13 and 3.14.

Next, we consider the case for $\overline{a}_0 \in (1/c_0, b_0)$, that is, $(\overline{a}_0, b_0, c_0) \in \mathcal{P}$. Then we see that the equilibrium points of (1.5a) are $(0,0)$, $(0, \overline{a}_0)$, $(1,0)$, and

$$\left(\frac{1 - \overline{a}_0\, c_0}{1 - b_0\, c_0}, \frac{\overline{a}_0 - b_0}{1 - b_0\, c_0} \right).$$

By (3.20), we have

$$0 < \frac{1 - \overline{a}_0\, c_0}{1 - b_0\, c_0} < \overline{u}_0(0) = \frac{1}{2} \left\{ \frac{1 - \overline{a}_0\, c_0}{1 - b_0\, c_0} + 1 \right\} < 1.$$

From Lemmas 3.13, 3.15, and 3.16, we obtain $\overline{\boldsymbol{u}}_0(-\infty) = (0, \overline{a}_0)$ and $\overline{\boldsymbol{u}}_0(+\infty) = (1,0)$. Then we have $(b_0, c_0) \in \operatorname{Int} \mathcal{E}_0(\beta_0)$ by virtue of $(\overline{a}_0, b_0, c_0) \in \mathcal{P}$ and Lemma 3.11. This contradicts the definition of (b_0, c_0).

Therefore we obtain $\mathcal{E}_0(\beta_0) = \mathcal{P}_0$. For each $(b,c) \in \mathcal{P}_0$, we define

$$a^*(b,c,\beta_0) = \sup\{\, a \mid [\overline{a}(b,c,\beta_0), a) \times \{\,(b,c)\,\} \subset \mathcal{E}\,\},$$
$$a_*(b,c,\beta_0) = \inf\{\, a \mid (a, \overline{a}(b,c,\beta_0)] \times \{\,(b,c)\,\} \subset \mathcal{E}\,\}.$$

By Lemma 3.12, we have

$$1/c \le a_*(b,c,\beta_0) < \overline{a}(b,c,\beta_0) < a^*(b,c,\beta) \le b$$

for any $(b,c) \in \mathcal{P}_0$. It follows from Lemma 3.5 that the functions $\boldsymbol{u}(.;a,b,c) = (u,v)(.;a,b,c)$ and $s(a,b,c)$ given in Lemma 3.12 can be regarded as C^1-class single-valued functions in $(a,b,c) \in \mathcal{E}$.

We assume $a^*(b_0,c_0,\beta_0) < b_0$ for some $(b_0,c_0) \in \mathcal{P}_0$. Then we have $(a^*(b_0,c_0,\beta_0), b_0, c_0) \in \mathcal{P}$. Let $\{\, a_n \,\}_{n=1}^{\infty} (\subset (\overline{a}(b_0,c_0,\beta_0), a^*(b_0,c_0,\beta_0)))$ be an arbitrary increasing sequence which satisfies $\lim_{n\to\infty} a_n = a^*(b_0,c_0,\beta_0)$. By Lemmas 3.6 and 3.12, we see that $\{\, s(a_n,b_0,c_0)\,\}_{n=1}^{\infty} (\subset (0, 2\sqrt{b_0 d})$ is a increasing sequence, i.e., the limit $\lim_{n\to\infty} s(a_n,b_0,c_0)\,(\equiv s_0 > 0)$ exists. Then we have

$$\| \boldsymbol{u}(.;a_n,b_0,c_0) \|_{C^2(\mathbf{R})} \le \frac{C_5(b_0,c_0)}{s(a_n,b_0,c_0)} \le \frac{C_5(b_0,c_0)}{s(a_1,b_0,c_0)}$$

for any $n \ge 1$ because of Lemma 3.6. By using an argument similar to the one above, it follows that, for any fixed $\xi_{14} > 0$, there exist $\{\, a_{n_j} \,\}_{j=1}^{\infty} (\subset \{\, a_n \,\}_{n=1}^{\infty})$ and $\boldsymbol{u}_0(\xi)$ such that $\boldsymbol{u}_0(\xi)$ is a strictly monotone solution of (1.5) for $(a,b,c,s) = (a^*(b_0,c_0,\beta_0), b_0, c_0, s_0)$ and satisfies

$$\| \boldsymbol{u}(.;a_{n_j},b_0,c_0) - \boldsymbol{u}_0 \|_{C^2([-\xi_{14},\xi_{14}])} \to 0$$

as $j \to \infty$. By Lemma 3.12, we have $(a^*(b_0,c_0,\beta_0), b_0, c_0) \in \operatorname{Int}\mathcal{E}$. This contradicts the definition of $a^*(b,c,\beta_0)$. Thus we obtain $a^*(b,c,\beta_0) = b$ for any $(b,c) \in \mathcal{P}_0$. In a similar manner, we can also show that $a_*(b,c,\beta_0) = 1/c$ holds for any $(b,c) \in \mathcal{P}_0$. Therefore, we have $\mathcal{E} = \mathcal{P}$.

Finally we shall show that $\overline{a}(b,c,\beta_0)$ is independent of β_0. Let $\overline{a}_1$ be an arbitrary constant such that (1.5) with $(a,s) = (\overline{a}_1, 0)$ has a strictly monotone solution for $(b,c) \in \mathcal{P}_0$. We assume that $\overline{a}(b,c,\beta_0) \ne \overline{a}_1$. Since $s(a,b,c)$ is defined on $\mathcal{P}$ and satisfies $s_a(a,b,c) > 0$, we have

$$0 = s(\overline{a}(b,c,\beta_0),b,c) \begin{cases} < s(\overline{a}_1,b,c) = 0 & \text{if } \overline{a}(b,c,\beta_0) < \overline{a}_1, \\ > s(\overline{a}_1,b,c) = 0 & \text{if } \overline{a}(b,c,\beta_0) > \overline{a}_1. \end{cases}$$

This contradiction implies that $\overline{a}(b,c,\beta_0) = \overline{a}_1$ holds. □

Acknowledgment. The author expresses his sincere gratitude to referees for valuable comments and encouragement.

REFERENCES

[1] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[2] C. CONLEY AND R. GARDNER, *An application of the generalized Morse index to travelling wave solutions of a competitive reaction-diffusion model*, Indiana Univ. Math. J., 33 (1984), pp. 319–343.

[3] W. A. COPPEL, *Dichotomies in Stability Theory*, Springer-Verlag, Berlin, 1978.

[4] R. A. GARDNER, *Existence and stability of travelling wave solutions of competition models: A degree theoretic approach*, J. Differential Equations, 44 (1982), pp. 343–364.

[5] Y. HOSONO, *Singular perturbation analysis of travelling waves for diffusive Lotka-Volterra competition models*, in Numerical and Applied Mathematics Part II, C. Brezinski, ed., Baltzer, Basel, 1989, pp. 687–692.

[6] Y. HOSONO AND M. MIMURA, *Singular perturbation approach to traveling waves in competing and diffusing species models*, J. Math. Kyoto Univ., 22 (1982), pp. 435–461.

[7] T. IKEDA AND M. MIMURA, *An interfacial approach to regional segregation of two competing species mediated by a predator*, J. Math. Bid., 31 (1993), pp. 215–240.

[8] Y. KAN-ON AND E. YANAGIDA, *Existence of non-constant stable equilibria in competition-diffusion equations*, Hiroshima Math. J., 23 (1993), pp. 193–221.

[9] H. KOKUBU, *Homoclinic and heteroclinic bifurcations of vector fields*, Japan J. Appl. Math., 5 (1988), pp. 455–501.

[10] M. MIMURA AND P. C. FIFE, *A 3-component system of competition and diffusion*, Hiroshima Math. J., 16 (1986), pp. 189–207.

[11] J. D. MURRAY, *Mathematical Biology*, Springer-Verlag, Berlin, New York, 1989.

[12] M. M. TANG AND P. C. FIFE, *Propagating fronts for competing species equations with diffusion*, Arch. Rational Mech. Anal., 73 (1980), pp. 69–77.

[13] A. I. VOL'PERT AND V. A. VOL'PERT, *Applications of the rotation theory of vector fields to the study of wave solutions of parabolic equations*, Trans. Moscow Math. Soc., 52 (1990), pp. 59–108.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 364–398, March 1995

006

THE CHILD–LANGMUIR LAW FOR THE BOLTZMANN EQUATION OF SEMICONDUCTORS*

NAOUFEL BEN ABDALLAH† AND PIERRE DEGOND†

Abstract. We investigate the so-called Child-Langmuir asymptotics of the one-dimensional stationary Boltzmann–Poisson system. The asymptotics apply when the lattice temperature is small and leads to a singular perturbation problem. We derive the limit problem associated with these asymptotics, and prove the existence of the Child-Langmuir current.

Key words. semiconductors, integral equation, Cauchy problem, contraction, nonlinear differential equation, nonlocal nonlinearity

AMS subject classifications. 34A12, 34A99, 45J05, 78A35, 82D99

1. Introduction.

The design of many high technology components in solid-state electronics, in vacuum diode technology or in high power hyperfrequency amplification requires an accurate description of charged-particle transport. Among all the possible models, the Vlasov or the Boltzmann equations, coupled with the Poisson or the Maxwell equations for the fields, provide the most accurate description of the physics of charged-particle transport. The numerical simulation of these models is an important tool for the designers.

The modeling of charged particle transport is particularly difficult, due to space-charge or internal boundary layers which very often appear. The description of such layers can usually be done by means of perturbation analyses of the stationary or time-dependent Vlasov–Poisson equations, which provide singular perturbation problems. One especially interesting problem was investigated by Langmuir and Compton [1], who showed that the charge boundary layer sitting at the cathode of a vacuum diode could produce a limitation of the current intensity which flows through the diode.

The mathematical analysis of this problem first started with a study of the boundary-value problem for the stationary Vlasov–Poisson equation in the one-dimensional cartesian case [2]. The perturbation problem and its convergence towards the reduced problem of [1] was analyzed in [3], in the same one-dimensional cartesian geometry. Then a numerical algorithm for the practical computation of the reduced solution was proposed in [4]. The passage to higher dimensions and more complicated models was investigated in [5] and [6], in which the well-posedness of the boundary value problems for the stationary Vlasov–Poisson, Vlasov–Maxwell, and Vlasov–Poisson–Boltzmann equations are proved in any dimension. Then by combining the ideas of [5], [6], and [3], the perturbation problem for the stationary cylindrically or spherically symmetric Vlasov–Poisson equation was investigated in [7].

Let us consider a simplified one-dimensional device which consists of two highly doped N^+ regions on each side of a lowly doped N^- region. Such an $N^+ - N^- - N^+$ device closely resembles a vacuum diode, where the metallic cathode and anode are replaced by the N^+ zones, and the vacuum region by the N^- zone. If the collisions

* Received by the editors April 2, 1993; accepted for publication (in revised form) September 14, 1993. This research was done while the first author was a member of the Centre de Mathématiques Appliquées of the Ecole Polytechnique, France, and the second author was a member of the Centre de Mathématiques et de Leurs Applications of the Ecole Normale Supérieure de Cachan, France.

† Mathématiques pour l'Industrie et la Physique Centre National de la Recherche Scientifique, Unité Mixte de Recherche 9974, UFR MIG, Université Paul Sabatier Toulouse 3, 118 route de Narbonne 31062 Toulouse Cedex, France.

of the carriers with the crystal lattice defects were negligible, we could use the same system of stationary Vlasov–Poisson equations and the perturbation problem would lead to the same result as in the vacuum diode case. This was remarked by Shur and Eastman in [8]. Strictly speaking, this asymptotic solution is only relevant for either large direct biases or low lattice temperatures.

However, the collisionless approximation is not valid in realistic situations. To investigate the effects of collisions, Shur and Eastman [9] proposed a model, based on a simplified one-dimensional hydrodynamic model and consisting of two equations of momentum and energy balance. In this paper, we show that the perturbation approach of the collisional kinetic model can be carried on. The reduced problem can be explicitly written, and the proof of its well-posedness is given under some restrictions. It reduces to the Langmuir and Compton [1] or Shur and Eastman [8] solution when the collision frequency vanishes. It also exhibits the same features; namely the current intensity cannot exceed a limiting value which depends on the collision frequency.

The outline of this paper is as follows: In the following section we derive the model which follows from formal asymptotics of the Vlasov–Poisson–Boltzmann system of semiconductors. Its solution is entirely determined by means of the electrostatic potential solution of a semilinear elliptic problem:

$$\varphi'' = n_1 + n_2, \quad \varphi(0) = 0, \quad \varphi(1) = 1, \tag{1}$$

where n_1 is an explicit nonlinear function of φ and n_2 is a solution of an integral equation depending on φ. In §3, we show that this integral equation has a unique solution. In the fourth section, we consider the Cauchy problem

$$\varphi'' = n_1 + n_2, \qquad \varphi(0) = 0, \quad \varphi'(0) = \beta, \tag{2}$$

and prove its well-posedness. Finally, in §5, we prove the existence and uniqueness of the Child-Langmuir current for a large range of values of the relaxation time. The Child-Langmuir current is, as usual, defined as the current for which the solution φ of the boundary value problem (1) has a vanishing derivative at $x = 0$.

2. The model. We consider a one-dimensional unipolar semiconductor structure, which consists of two highly doped N^+ regions on each side of a lowly doped N^- region. In such a structure, the N^+ regions behave approximately like metallic contacts and the N^- region can be modeled just like a vacuum diode, by assuming that the injection of carriers at the $N^+ - N^-$ junctions can be described by a given emission profile $G(V)$ on the source side, and can be neglected on the drain side (see [10] for more details). Of course, this model is very crude (see [11] for more realistic models), but it focuses the analysis on the injection process, which is of primary importance for the global behavior of many devices.

We assume that the $N^+ - N^-$ junctions are located at $X = 0$ (for the source side), and $X = L$ (for the drain side), and that the N^- region is represented by the interval $[0, L]$. The electron distribution function $F(X, V)$, the electric potential $\Phi(X)$, and the electron concentration $N(X)$ satisfy the system of stationary Vlasov–Poisson–Boltzmann equations of semiconductors:

$$V\frac{\partial F}{\partial X} + \frac{e}{m}\frac{d\Phi}{dX}\frac{\partial F}{\partial V} = Q(F), \qquad X \in [0, L], \quad V \in \mathbb{R} \tag{3}$$

$$\frac{d^2\Phi}{dX^2} = \frac{e}{\varepsilon_0} N(X), \quad X \in [0, L], \tag{4}$$

$$N(X) = \int_{-\infty}^{+\infty} F(X,V)\,dV, \qquad X \in [0,L]. \tag{5}$$

We denote the collision operator by $Q(F)$, which models the interaction of the electrons with the crystal defects. The reader will find in [12]–[14] a fairly complete description of these interactions. In this paper, we will restrict our analysis to the relaxation time model

$$Q(F) = -\frac{1}{\mathcal{T}}\left(F(X,V) - N(X)M_T(V)\right), \tag{6}$$

where $\mathcal{T} > 0$ is the relaxation time, $M_T(V)$ is the normalized Maxwellian distribution associated with the lattice temperature T

$$M_T(V) = \frac{1}{\sqrt{\frac{2\pi k_B T}{m}}} \exp\left(-\frac{mV^2}{2k_B T}\right), \tag{7}$$

and k_B is the Boltzmann constant. Finally, ε_0 denotes the medium permittivity. In the Poisson equation (4), we neglect the doping density of the N^- region, which is a fairly good assumption (see [15]). However, the analysis can also be performed with a nonvanishing N^- in the collisionless case (see [20]).

The system (3)–(5) is supplemented with the following boundary conditions (see [10] for details):

$$F(0,V) = G(V), \qquad V > 0, \tag{8}$$

$$F(L,V) = 0, \qquad V < 0, \tag{9}$$

$$\Phi(0) = 0, \qquad \Phi(L) = \Phi_L > 0. \tag{10}$$

Indeed, Φ_L is the applied bias, and since the potential is nearly constant in the N^+ regions, it applies entirely at the boundary of the N^- region. Since the N^+ region on the source side is close to a state of thermal equilibrium with the crystal lattice, it is natural to assume that the injection profile is given by

$$G(V) = N^+ M_T(V), \qquad V > 0, \tag{11}$$

where N^+ is the doping density of the N^+ region. On the other side, the injection is negligible, which implies (9). More generally, we shall assume that $G(V)$ is a given function such that

$$\int_0^\infty G(V)dV = \frac{N^+}{2}$$

and such that the thermal emission velocity V_G given by

$$V_G = \left(\int_0^\infty V^2 G(V)\,dV \Big/ \int_0^\infty G(V)\,dV\right)^{\frac{1}{2}} \tag{12}$$

is equal to the thermal velocity associated with the maxwellian (7)

$$V_G = V_{th} = \sqrt{\frac{k_B T}{m}}.$$

Because of this, we assume that the lattice temperatures in the N^+ and N^- region are the same, which is a fairly unrestrictive assumption for physical applications.

The existence of solutions of the system (3)–(5), (8)–(10) is mathematically proven in [5]. Its numerical solution has been achieved by iterative methods in [15] and [16], by particle methods in [17] and [18], and by Monte-Carlo methods (see [12] and the references therein).

In the "Child-Langmuir" regime, the thermal emission energy is small compared with the applied bias, while the injected current remains finite [3]:

$$\frac{mV_G^2}{2} \ll e\Phi_L$$

or

$$V_G \ll V_L, \qquad V_L = \sqrt{\frac{2e\Phi_L}{m}}. \tag{13}$$

Therefore, we introduce a "small" parameter ε by

$$\varepsilon = \frac{V_G}{V_L} \ll 1. \tag{14}$$

We shall use L, V_L, and Φ_L as characteristic length, velocity, and potential scales. We introduce auxiliary units of density $\overline{N}$, current density $\overline{J}$, distribution function $\overline{F}$, and relaxation time $\overline{T}$, according to

$$\overline{N} = \frac{\varepsilon_0 \Phi_L}{eL^2}, \qquad \overline{J} = e\overline{N}V_L, \qquad \overline{F} = \frac{\overline{N}}{V_L}, \qquad \overline{T} = \frac{L}{V_L}, \tag{15}$$

and use the following scaling:

$$\begin{cases} X = Lx, & V = V_L v, & \Phi = \Phi_L \varphi, \\ N = \overline{N} n, & J = -\overline{J} j, & F = \overline{F} f, \\ & \mathcal{T} = \overline{T}\tau. \end{cases} \tag{16}$$

Furthermore, we introduce a dimensionless profile $g(v)$ and express $G(V)$ according to

$$\overline{F}^{-1} G(V) = \frac{1}{\varepsilon^2} g\left(\frac{V}{V_G}\right) = \frac{1}{\varepsilon^2} g\left(\frac{v}{\varepsilon}\right). \tag{17}$$

The expression (17) means that V_G is the characteristic velocity associated with G and is small, while the factor ε^2 insures that the injected current J_G,

$$J_G = \int_0^\infty VG(V)dV,$$

remains independent of ε in the units of $\overline{J}$. We recall that these two facts are the key hypotheses of the Child-Langmuir asymptotics [3]. We also note that a different normalizing condition of the injection profile G can contribute to a finite built-in potential at the $N^+ - N^-$ interfaces (see [19]). Let us introduce

$$g^\varepsilon(v) = \frac{1}{\varepsilon^2} g\left(\frac{v}{\varepsilon}\right) \tag{18}$$

and write the scaled Vlasov–Poisson–Boltzmann equation

$$v\frac{\partial f^\varepsilon}{\partial x} + \frac{1}{2}\frac{d\varphi^\varepsilon}{dx}\frac{\partial f^\varepsilon}{\partial v} = -\frac{1}{\tau}\left(f^\varepsilon - n^\varepsilon \frac{1}{\varepsilon}M_0\left(\frac{v}{\varepsilon}\right)\right), \tag{19}$$

$$\frac{d^2\varphi^\varepsilon}{dx^2} = n^\varepsilon(x), \quad x \in [0,1], \tag{20}$$

$$n^\varepsilon(x) = \int_{-\infty}^{+\infty} f^\varepsilon(x,v)\,dv, \quad x \in [0,1], \tag{21}$$

$$f^\varepsilon(0,v) = g^\varepsilon(v), \quad v > 0, \tag{22}$$

$$f^\varepsilon(1,v) = 0, \quad v < 0, \tag{23}$$

$$\varphi^\varepsilon(0) = 0, \quad \varphi^\varepsilon(1) = 1, \tag{24}$$

where

$$M_0(v) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{v^2}{2}\right). \tag{25}$$

Our aim is to solve the reduced problem corresponding to $\varepsilon = 0$. So we pass formally to the limit in the equations and get the following reduced problem:

$$v\frac{\partial f}{\partial x} + \frac{1}{2}\frac{d\varphi}{dx}\frac{\partial f}{\partial v} = -\frac{1}{\tau}(f - n\,\delta(v)), \tag{26}$$

$$\frac{d^2\varphi}{dx^2} = n(x), \quad x \in [0,1], \tag{27}$$

$$n(x) = \int_{-\infty}^{+\infty} f(x,v)\,dv, \quad x \in [0,1]. \tag{28}$$

$$f(1,v) = 0, \quad v < 0, \tag{29}$$

$$\varphi(0) = 0, \quad \varphi(1) = 1, \tag{30}$$

where $\delta(v)$ is the delta function; the formal limit of (22) can be expressed by

$$\text{supp}\,\{f(0,v), v \geq 0\} = \{v = 0\}, \qquad 0 < \int_0^\infty v f(0,v)\,dv < \infty. \tag{31}$$

In the collisionless case ($\tau = \infty$, see [3]) the condition (31) forces the solution to be a positive measure supported by the characteristics issued from the point $(x,v) = (0,0)$. In the present case ($\tau < \infty$), it is natural to think that the solution will exhibit the same features. This means that the particles fly from the cathode to the anode,

following the same characteristics. However some particles will undergo a collision in their way to the anode. The collision operator at the right-hand side of (26) sends the velocity of the colliding particles back to zero. Thus, after the collision, the particle follows another characteristic, namely the one issued from the point $(y, 0)$ where y is the location of the collision. This new characteristic is given by the equation

$$v^2 - \varphi(x) = \text{constant} = -\varphi(y),$$

that is,

$$v = \sqrt{\varphi(x) - \varphi(y)}. \tag{32}$$

The + sign has to be retained because no electron is emitted at the anode, and then the particles flow from the cathode to the anode. Of course, once the particle is on the secondary characteristic (32), it can suffer a second collision which sends it to a third characteristic and so on. Therefore, in addition to a positive measure supported by the principal characteristics (i.e., issued from $(0,0)$), the solution $f(x,v)$ contains a superposition of the contributions of each of the secondary characteristics (32) for $y \in [0.1]$. Therefore we write it as

$$f(x,v) = n_1(x)\,\delta(v - \sqrt{\varphi(x)}) + \int_0^x \overline{n}_2(x,y)\delta(v - \sqrt{\varphi(x) - \varphi(y)})\,dy, \tag{33}$$

where $n_1(x)$ is the density of particles carried by the principal characteristics and $\overline{n}_2(x,y)\,dy$ is the density of particles carried by the bunch of characteristics issued between the point $(y,0)$ and $(y+dy,0)$. Mathematically speaking, the integral on the right-hand side of (33) can be viewed as a change of variables.

We introduce

$$j_1(x) = n_1(x)\sqrt{\varphi(x)}, \tag{34}$$

$$\overline{j}_2(x,y) = \overline{n}_2(x,y)\,\sqrt{\varphi(x) - \varphi(y)}, \tag{35}$$

$$n_2(x) = \int_0^x \overline{n}_2(x,y)\,dy, \tag{36}$$

$$j_2(x) = \int_0^x \overline{j}_2(x,y)\,dy = \int_0^x \overline{n}_2(x,y)\,\sqrt{\varphi(x) - \varphi(y)}\,dy. \tag{37}$$

j_1 is the current flowing along the principal chracteristics, and n_2 and j_2 are the density and current carried by all the secondary characteristics. We have

$$n_1(x) + n_2(x) = n(x), \tag{38}$$

$$j_1(x) + j_2(x) = j = \text{constant}. \tag{39}$$

Inserting the expression (33) into the equation (26), we formally find the following equation:

$$\left(\frac{dj_1}{dx} + \frac{n_1(x)}{\tau}\right)\delta(v - \sqrt{\varphi(x)}\,)$$
$$+ \int_0^x \left(\frac{\partial \overline{j}_2}{\partial x}(x,y) + \frac{\overline{n}_2(x,y)}{\tau}\right)\delta(v - \sqrt{\varphi(x) - \varphi(y)}\,)\,dy$$
$$= \frac{1}{\tau}\,n(x)\,\delta(v) = \int_0^x \frac{n(y)}{\tau}\,\delta(y - x)\delta(v - \sqrt{\varphi(x) - \varphi(y)}\,)\,dy.$$

The terms in factor of $\delta(v - \sqrt{\varphi(x)})$ and of $\delta(v - \sqrt{\varphi(x) - \varphi(y)})$ can be separated, which leads to

$$\frac{dj_1}{dx} + \frac{n_1(x)}{\tau} = 0, \tag{40}$$

$$\frac{\partial \bar{j}_2}{\partial x}(x,y) + \frac{\bar{n}_2(x,y)}{\tau} = 0, \quad 0 < y < x < 1, \tag{41}$$

$$\bar{j}_2(y,y) = \frac{1}{\tau} n(y). \tag{42}$$

Equation (40) means that the current decreases as one moves along the principal characteristics, because of the collisions. The same is true for each one of the secondary characteristics (equation (41)). Finally, equation (42) specifies that the current carried by one of the secondary characteristics at its starting point y is made of the contribution of all the particles which have collided at this point, either from the principal or the secondary characteristics (cf. (38)). Now we must specify the initial condition for equation (40). For this, it is natural to assume that the current carried by the secondary characteristics vanishes for $x = 0$,

$$\lim_{x \to 0} j_2(x) = 0. \tag{43}$$

Indeed, for small values of x the collisions are negligible, the total current is carried by the principal characteristics, then we deduce from (39) that

$$j_1(0) = j. \tag{44}$$

The total current j that flows through the device will be assumed arbitrary for the moment. By keeping in mind the relations (34) and (35), we can see that the solutions of the equations (40)–(42) and (44) can be written explicitly:

$$j_1(x) = j \exp\Big(-\frac{1}{\tau} \int_0^x \frac{dz}{\sqrt{\varphi(z)}} \Big), \tag{45}$$

$$\bar{j}_2(x,y) = \frac{n(y)}{\tau} \exp\Big(-\frac{1}{\tau} \int_y^x \frac{dz}{\sqrt{\varphi(z) - \varphi(y)}} \Big). \tag{46}$$

For a given φ, equation (45) gives a closed expression of j_1 and (thanks to (34)) the density n_1,

$$n_1(x) = \frac{j}{\sqrt{\varphi(x)}} \exp\Big(-\frac{1}{\tau} \int_0^x \frac{dz}{\sqrt{\varphi(z)}} \Big). \tag{47}$$

The situation is different for $\bar{j}_2$, since $\bar{j}_2$ depends on n, which in turn depends on $\bar{j}_2$ via equations (36) and (35). From now on, we let

$$g_\varphi(x,y) = \exp\Big(-\frac{1}{\tau} \int_y^x \frac{dz}{\sqrt{\varphi(z) - \varphi(y)}} \Big). \tag{48}$$

By using (36), (35), and (46), we obtain

$$n_2(x) - \int_0^x \frac{g_\varphi(x,y)}{\tau\sqrt{\varphi(x)-\varphi(y)}} n_2(y)\,dy = \int_0^x \frac{g_\varphi(x,y)}{\tau\sqrt{\varphi(x)-\varphi(y)}} n_1(y)\,dy. \tag{49}$$

Since n_1 has been previously determined by (47), equation (49) is an integral equation for n_2. Then we can write the complete system satisfied by the potential

$$\frac{d^2\varphi}{dx^2} = n(x), \tag{50}$$

$$\varphi(0) = 0, \qquad \varphi(1) = 1, \tag{51}$$

$$n(x) = n_1(x) + n_2(x), \tag{52}$$

where n_1 is given by (47) and n_2 is the solution of (49), and where j in (47) is an arbitrary nonnegative constant. The following results are inspired from the existence of the Child-Langmuir current in the collisionless case (the vacuum diode [3]) and are proven in this paper.

THEOREM 2.1. *There exist* $\tau_1 \geq 7/9$ *and* $\tau_2 \leq 4/5$ *such that for evry* $\tau \in]0,\tau_1] \bigcup [\tau_2, \infty[$, *there exists a unique value* $j = j_{CL}(\tau)$ *such that the problem* (50)–(52), (47), (49) *has a unique solution* φ *that satisfies* $d\varphi/dx(0) = 0$. *Moreover,* $j_{CL}(\tau) \sim 4/9$ *when* τ *tends to* ∞, *and* $j_{CL}(\tau) \sim \tau 9/16$ *when* τ *tends to zero.*

This result expressed in the physical variables gives the expression of the Child-Langmuir current

$$J_{CL} = -j_{CL}\left(\frac{\mathcal{T}}{L}\sqrt{\frac{2e\Phi_L}{m}}\right)\varepsilon_0\sqrt{\frac{2e}{m}}\,\frac{\Phi_L^{3/2}}{L^2}.$$

In the collisionless limit ($\tau \to \infty$), we find the Child-Langmuir current of the vacuum diode [3], and in the collision-dominated limit ($\tau \to 0$), we find

$$J_{CL} = -\frac{9}{16}\mathcal{T}\varepsilon_0\left(\frac{2e}{m}\right)\frac{\Phi_L^2}{L^3},$$

or in terms of the mobility $\mu = \frac{e\mathcal{T}}{m}$ and of the external field $E_L = \frac{\Phi_L}{L}$,

$$J_{CL} = -\frac{9}{8}\varepsilon_0\mu\frac{\Phi_L^2}{L^3} = -\frac{9}{8}\varepsilon_0\mu\frac{E_L^2}{L}.$$

It is remarkable that this formula completely differs from the one used for a homogeneous semiconductor,

$$J = e\mu n E_L,$$

where n is the density of free carriers. Here the only available free carriers are those provided by the injection contact and lead to a current which is proportional to the squared external electric field. As far as we know, such a formula has not been found

previously in the physical literature. (See [21] for an overview of the current-voltage characteristics of the basic semiconductor devices.)

In the vacuum diode case, it is proved that the limit current cannot exceed the Child-Langmuir current. In our case, we prove that a limitation of the limit current occurs.

THEOREM 2.2. *There exists a value* $j_{\max}(\tau)$ *such that the system* (50)–(52), (47), (49) *has no solution for* $j > j_{\max}(\tau)$. *This value satisfies the following estimate:*

$$j_{CL}(\tau) \leq j_{\max}(\tau) < \min\left(\frac{4}{9}, \frac{9}{16}\tau\right).$$

It has been proved neither that the problem (50)–(52), (47), (49) has a unique solution for $j < j_{\max}(\tau)$, nor that $j_{\max}(\tau) = j_{CL}(\tau)$, as could be conjectured from the inspection of the vacuum diode case [3] (however, $j_{\max}(\tau) \sim j_{CL}(\tau)$ in the vicinity of zero and infinity). Some pathologies, which occur in the vacuum diode case in higher dimensions [7], should make us careful about conjectures in this direction. Another open problem is the convergence of the solutions of the perturbed problem (19)–(24) to those of (33) and (47)–(52).

3. The reduced problem. To study the limit problem, we use a change of variable and unknowns in order to get rid of the constants j and τ. Thus, we define the constants λ and δ depending on j and τ according to

$$\delta = \frac{1}{\tau^3 j}, \quad \lambda = \tau^4 j^2 \tag{53}$$

and

$$\varphi(x) = \lambda\tilde{\varphi}(\delta x), \qquad n_1(x) = \lambda\mu^2\tilde{n}_1(\delta x), \quad n_2(x) = \lambda\mu^2\tilde{n}_2(\delta x). \tag{54}$$

With this rescaling, the system becomes, omitting the tildes,

$$\frac{d^2\varphi}{dx^2} = n(x), \tag{55}$$

$$n(x) = n_1(x) + n_2(x), \tag{56}$$

$$n_1(x) = \frac{1}{\sqrt{\varphi(x)}}\, g_\varphi(x, 0), \tag{57}$$

$$\begin{aligned} n_2(x) - \int_0^x \frac{g_\varphi(x,y)}{\sqrt{\varphi(x)-\varphi(y)}} n_2(y)\, dy \\ = \int_0^x \frac{g_\varphi(x,y)}{\sqrt{\varphi(x)-\varphi(y)}} n_1(y)\, dy, \end{aligned} \tag{58}$$

$$\varphi(0) = 0, \qquad \varphi(\delta) = \frac{1}{\lambda}, \tag{59}$$

$$g_\varphi(x,y) = \exp\left(-\int_y^x \frac{dz}{\sqrt{\varphi(z)-\varphi(y)}}\right). \tag{60}$$

The first step for the study of the above system consists in solving the integral equation (58) for a given electric potential φ. This will be done in the following subsection, but we first begin with the following definition.

DEFINITION 3.1. *Let $\alpha \in [1,2)$, $\mu > 0$ and $C > 0$ be given. We define $Q(\alpha, \mu, C)$ as the set of convex functions φ on $[0,\mu]$ such that*

$$\varphi(x) > 0, \quad x \in (0,\mu],$$

$$\varphi(0) = 0, \quad \text{and } |\varphi(x) - \varphi(y)| \geq C\,|x-y|^{\alpha}, \quad x, y \in [0,\mu], \tag{61}$$

and we set

$$L^{\infty}_{\alpha}(0,\mu) = L^{\infty}(0,\mu, x^{\alpha-1}dx) = \{\psi,\, x^{\alpha-1}\,\psi(x) \in L^{\infty}(0,\mu)\}.$$

3.1. The integral equation. Here we consider a given convex nonnegative function and we show the existence and uniqueness of a solution of (58).

PROPOSITION 3.2. *Let φ be given in $Q(\alpha,\mu,C)$ for some α in $[1,2)$, μ and C positive. Then the equation (58) has a unique solution n_2 in $L^{\infty}_{\alpha}(0,\mu)$. This solution has the following properties:*

(1) *n_2 is positive for $x \in (0,\mu)$.*

(2) *The norm of n_2 in $L^{\infty}_{\alpha}(0,\mu)$ can be bounded by a constant depending only on C, α, and μ.*

To prove this proposition we define the map

$$\begin{aligned} K_{\varphi} : L^{\infty}_{\alpha}(0,\mu) &\longrightarrow L^{\infty}_{\alpha}(0,\mu), \\ m(x) &\longmapsto K_{\varphi} m, \end{aligned} \tag{62}$$

such that

$$K_{\varphi} m(x) = G(x) + \int_0^x \frac{g_{\varphi}(x,y)}{\sqrt{\varphi(x) - \varphi(y)}} m(y)\, dy, \tag{63}$$

where

$$G(x) = \int_0^x \frac{g_{\varphi}(x,y)}{\sqrt{\varphi(x) - \varphi(y)}} n_1(y)\, dy. \tag{64}$$

We have the following lemma.

LEMMA 3.3. *The map K_{φ} takes $L^{\infty}_{\alpha}(0,\mu)$ into $L^{\infty}_{\alpha}(0,\mu)$. Moreover , for all m_1 and m_2 in $L^{\infty}_{\alpha}(0,\mu)$ and p integer, the following estimate holds on $[0,\mu]$:*

$$x^{\alpha-1}\,|K^p_{\varphi} m_1(x) - K^p_{\varphi} m_2(x)| \leq C_1^p\, x^{p(1-\frac{\alpha}{2})}\, I_1\, I_2 \cdots I_p \|m_1 - m_2\|_{L^{\infty}_{\alpha}(0,\mu)}, \tag{65}$$

where

$$I_k = \int_0^1 \frac{t^{k\,(1-\frac{\alpha}{2}) - \frac{\alpha}{2}}}{(1-t)^{\frac{\alpha}{2}}}\, dt, \tag{66}$$

and the constant $C_1 = \frac{1}{\sqrt{C}}$.

Proof. Since the definition of $Q(\alpha, \mu, C)$ implies that $\varphi(x) \geq C\, x^{\alpha}$, we deduce from (57) that

$$n_1(x) \leq \frac{1}{\sqrt{C}\, x^{\frac{\alpha}{2}}},$$

and then we have

$$n_1 \in L^{\infty}_{\alpha}(0, \mu).$$

Thus, to prove that $K_{\varphi}(L^{\infty}_{\alpha}) \subset L^{\infty}_{\alpha}$, it is sufficient to prove that

$$h(x) = \int_0^x \frac{g_{\varphi}(x, y)}{\sqrt{\varphi(x) - \varphi(y)}} m(y)\, dy \tag{67}$$

is in L^{∞}_{α} for every m in L^{∞}_{α}.

First, we have

$$x^{\alpha-1}|h(x)| \leq x^{\alpha-1} \int_0^x \frac{\|m\|_{L^{\infty}_{\alpha}(0,\mu)}\, dy}{\sqrt{C}\, y^{\alpha-1}\, (x-y)^{\frac{\alpha}{2}}}.$$

The change of variable $y = xt$ in this integral gives

$$x^{\alpha-1}|h(x)| \leq \frac{x^{1-\frac{\alpha}{2}}}{\sqrt{C}}\, I_1 \|m\|_{L^{\infty}_{\alpha}(0,\mu)} \tag{68}$$

and ensures that $h \in L^{\infty}_{\alpha}(0, \mu)$ since $\alpha < 2$.

Now we prove the estimate (65). Let us set $m = m_1 - m_2$ and $h = K_{\varphi} m_1 - K_{\varphi} m_2$. Then m and h satisfy (67) and the application of (68) proves (65) for $p = 1$. To prove the estimate for all p, we proceed by induction. (The details are left to the reader.) □

LEMMA 3.4. *There exist two positive constants C_2 and γ depending only on α, such that for every k,*

$$I_k \leq \frac{C_2}{k^{\gamma}}.$$

Proof. By the Hölder inequality we find

$$\begin{aligned} I_k &= \int_0^1 \frac{t^{k\,(1-\frac{\alpha}{2})\,-\,\frac{\alpha}{2}}}{(1-t)^{\frac{\alpha}{2}}}\, dt \\ &\leq \left(\int_0^1 t^{k\,(1-\frac{\alpha}{2})\,\beta'}\, dt \right)^{\frac{1}{\beta'}} \left(\int_0^1 \frac{dt}{t^{\frac{\alpha\beta}{2}} (1-t)^{\frac{\alpha\beta}{2}}} \right)^{\frac{1}{\beta}}, \end{aligned}$$

where β is chosen such that $\frac{\alpha\beta}{2} < 1$. Thus, there exists a constant K such that

$$I_k \leq \frac{K}{(1 + k\,(1 - \frac{\alpha}{2})\,\beta')^{\frac{1}{\beta'}}} \leq \frac{C_2}{k^{\frac{1}{\beta'}}}. \tag{69}$$

□

End of the proof of Proposition 3.2. Lemma 3.3, together with Lemma 3.4, implies

$$\|K^p_{\varphi} m_1 - K^p_{\varphi} m_2\|_{L^{\infty}_{\alpha}(0,\mu)} \leq \frac{\|m_1 - m_2\|_{L^{\infty}_{\alpha}(0,\mu)}}{(p!)^{\gamma}} \left[\frac{C_2 \mu^{1-\frac{\alpha}{2}}}{\sqrt{C}} \right]^p,$$

which ensures that K_φ^p is a strict contraction for p large enough. Thus, K_φ has a unique fixed point and this fixed point is the limit of all the sequences

$$n_2^k = K_\varphi(n_2^{k-1}). \tag{70}$$

The positivity of n_2 comes from the positiveness of n_1. Indeed, by induction it is easy to prove that the sequence n_2^k, defined by formula (70) and such that $n_2^1 = 0$, is always positive. Finally, it is readily seen that all the estimates established in this section depend on φ by means of α, μ, and C only.

4. The Cauchy problem. In order to prove the existence of a solution of the problem (55)–(60), we proceed as in [3] or [7] and introduce the Cauchy problem

$$\begin{cases} \dfrac{d^2\varphi}{dx^2} = n(x), \\ n(x) = n_1(x) + n_2(x), \\ \varphi(0) = 0, \quad \dfrac{d\varphi}{dx}(0) = \beta \geq 0, \end{cases} \tag{71}$$

where n_1 and n_2 are given by the formulae (57) and (58). We now give the main theorem of this section.

THEOREM 4.1. *For every fixed value $\beta \geq 0$ there exists a unique solution φ of the Cauchy problem* (71) *defined on $\mathbb{R}^+$. Moreover, this solution is equivalent near $x = 0$ to the solution of*

$$\frac{d^2 f}{dx^2} = \frac{1}{\sqrt{f}}, \qquad f(0) = 0, \quad \frac{df}{dx}(0) = \beta.$$

By "equivalent" we mean that the ratio of the two functions goes to one as x goes to zero.

The proof of this theorem is quite long and will be divided in two steps: First we will prove the existence and uniqueness of the solution on a small interval near zero. Then we will prove the existence and uniqueness on any bounded interval of $\mathbb{R}^+$. The first step will itself contain two cases: The case $\beta = 0$ which is more singular than the case $\beta > 0$. (See also [7], part 2 for similarities with the present problem.)

In the following two subsections we show the local existence and uniqueness of solutions; this result is stated in the next proposition.

PROPOSITION 4.2. *For every fixed value $\beta \geq 0$, there exists a positive constant μ such that the Cauchy problem* (71) *has a unique solution φ on $[0, \mu]$. Moreover, this solution is equivalent near $x = 0$ to the solution of*

$$\frac{d^2 f}{dx^2} = \frac{1}{\sqrt{f}}, \qquad f(0) = 0, \quad \frac{df}{dx}(0) = \beta. \tag{72}$$

Remark. Although the Cauchy problem is nonlinear and nonlocal, we can treat it as a differential equation because $n(x)$ only depends on $\varphi(y)$ for $y \leq x$. This is also important for the numerical computation of φ.

Remark. The solution of equation (72) is given implicitly by

$$\int_0^{f(x)} \frac{dg}{\sqrt{\beta^2 + 4\sqrt{g}}} = x, \quad x \in [0. + \infty);$$

in particular, for $\beta = 0$, we have $f(x) = \left(\frac{3x}{2}\right)^{4/3}$.

4.1. Local existence and uniqueness in the case $\beta > 0$. We begin the proof by showing some a priori estimates. For this aim, we first note that the solution of (71) (if it exists) is strictly convex. Then it satisfies

$$\varphi(x) \geq \beta x. \tag{73}$$

Thus $\varphi \in Q(1, \mu, \beta)$ for all $\mu > 0$, and then we have

$$n_1(x) \leq \frac{1}{\sqrt{\beta\, x}}. \tag{74}$$

By Theorem 3.2, $n_2 \in L^\infty(0, \mu)$ for every μ. Again by Theorem 3.2, we have

$$\|n_2\|_{L^\infty(0,\mu)} \leq K,$$

where K is a constant depending only on β and μ since $\varphi \in Q(1, \mu, \beta)$. Without any loss of generality, we take $\mu \leq 1$, and then the constant K can be chosen independently of μ. Combining the above estimates gives

$$0 \leq \frac{d^2\varphi}{dx^2}(x) \leq K + \frac{1}{\sqrt{\beta\, x}}.$$

Now choosing μ small enough, we get

$$0 \leq \frac{d^2\varphi}{dx^2}(x) \leq \frac{C_2}{\sqrt{x}} \qquad \text{on } (0, \mu).$$

By integrating this inequality twice we get

$$\left| \frac{d\varphi}{dx}(x) - \beta \right| \leq 2\, C_2 \sqrt{x}, \tag{75}$$

$$|\varphi(x) - \beta x| \leq \frac{4}{3}\, C_2\, x^{3/2}. \tag{76}$$

Now we can give the following definition.

DEFINITION 4.3. *We define the set $\mathcal{E}(\mu, \beta)$ as the set of nonnegative C^1 convex functions φ defined on $[0, \mu]$ such that*

$$\varphi(0) = 0, \quad \frac{d\varphi}{dx}(0) = \beta \tag{77}$$

and

$$\frac{1}{\sqrt{x}} \left| \frac{d\varphi}{dx}(x) - \beta \right| \in L^\infty(0, \mu). \tag{78}$$

Then, we define on $\mathcal{E}(\mu, \beta)$ the distance

$$d(\varphi_1, \varphi_2) = \sup_{x \in (0,\mu]} \frac{1}{\sqrt{x}} \left| \frac{d\varphi_1}{dx}(x) - \frac{d\varphi_2}{dx}(x) \right|. \tag{79}$$

Finally we define the map

$$\begin{aligned} \mathcal{S} :\ \mathcal{E}(\mu, \beta) &\to \mathcal{E}(\mu, \beta), \\ \varphi &\longmapsto \mathcal{S}(\varphi), \end{aligned} \tag{80}$$

such that

$$\text{(81)}\qquad \begin{cases} \dfrac{d^2\mathcal{S}(\varphi)}{dx^2} = n(\varphi)(x), \\ n(x) = n_1(\varphi)(x) + n_2(\varphi)(x), \\ \mathcal{S}(\varphi)(0) = 0, \qquad \dfrac{d\mathcal{S}(\varphi)}{dx}(0) = \beta \geq 0, \end{cases}$$

where $n_1(\varphi)$ and $n_2(\varphi)$ are defined by the formulae (57) *and* (58).

We begin with the following lemma.

LEMMA 4.4. *The set $\mathcal{E}(\mu,\beta)$ equipped with the distance d is a complete metric space.*

Proof. The proof is immediate and left to the reader. □

PROPOSITION 4.5. *There exist $\mu > 0$ and $k < 1$ (depending on β) such that for every function φ, ψ in $\mathcal{E}(\mu,\beta)$ the following estimates hold on $(0,\mu]$:*

$$\text{(82)}\qquad |n_1(\varphi)(x) - n_1(\psi)(x)| \leq \frac{k}{4\sqrt{x}}\, d(\varphi,\psi),$$

$$\text{(83)}\qquad |n_2(\varphi)(x) - n_2(\psi)(x)| \leq \frac{k}{4\sqrt{x}}\, d(\varphi,\psi).$$

Proof. We begin by proving (82). We have

$$\begin{aligned} n_1(\varphi)(x) - n_1(\psi)(x) = {} & \frac{1}{\sqrt{\varphi(x)}}\,[g_\varphi(x,0) - g_\psi(x,0)] \\ & + g_\psi(x,0)\left[\frac{1}{\sqrt{\varphi(x)}} - \frac{1}{\sqrt{\psi(x)}}\right]. \end{aligned} \qquad \text{(84)}$$

The second term of the right-hand side can be estimated as follows:

$$\begin{aligned} g_\psi(x,0)\left|\frac{1}{\sqrt{\varphi(x)}} - \frac{1}{\sqrt{\psi(x)}}\right| &\leq \frac{|\varphi(x) - \psi(x)|}{\sqrt{\varphi(x)}\sqrt{\psi(x)}\left(\sqrt{\varphi(x)} + \sqrt{\psi(x)}\right)} \\ &\leq \frac{1}{2\,\beta^{3/2}x^{3/2}} \int_0^x \left|\frac{d\varphi}{dx}(y) - \frac{d\psi}{dx}(y)\right| dy. \end{aligned} \qquad \text{(85)}$$

Using (79), the last estimate turns to

$$\text{(86)}\qquad g_\psi(x,0)\left|\frac{1}{\sqrt{\varphi(x)}} - \frac{1}{\sqrt{\psi(x)}}\right| \leq \frac{1}{3\,\beta^{3/2}}\, d(\varphi,\psi).$$

For the first term of (84) we have

$$\begin{aligned} |g_\varphi(x,0) - g_\psi(x,0)| &= \left|\exp\left(-\int_0^x \frac{dz}{\sqrt{\varphi(z)}}\right) - \exp\left(-\int_0^x \frac{dz}{\sqrt{\psi(z)}}\right)\right| \\ &\leq \left|\int_0^x \frac{dz}{\sqrt{\varphi(z)}} - \int_0^x \frac{dz}{\sqrt{\psi(z)}}\right| \\ &\leq \frac{x}{3\beta^{3/2}}\, d(\varphi,\psi). \end{aligned} \qquad \text{(87)}$$

Therefore we have

$$|n_1(\varphi)(x) - n_1(\psi)(x)| \leq C_3(1 + \sqrt{x})\, d(\varphi, \psi), \tag{88}$$

with a constant C_3 depending only on β. Now by choosing μ small enough so that $C_3(1 + \sqrt{x}) \leq \frac{k}{4\sqrt{x}}$ on $(0, \mu]$, we get the estimate (82). □

To prove (83), we proceed in a likewise manner, but the calculations are much more complicated since n_2 is not given explicitly. We consider two sequences, $n_2^p(\varphi)$ and $n_2^p(\psi)$, defined by

$$n_2^p(\varphi) = K_\varphi(n_2^{p-1}(\varphi)), \quad n_2^0(\varphi) = 0, \tag{89}$$

where K_φ is defined by formula (70) and

$$n_2^p(\psi) = K_\psi(n_2^{p-1}(\psi)), \quad n_2^0(\psi) = 0, \tag{90}$$

where K_ψ is defined analogously. Therefore, we have the following lemma.

LEMMA 4.6. *There exists a constant $C > 0$ which only depends on β such that*

$$\left|n_2^1(\varphi)(x) - n_2^1(\psi)(x)\right| \leq C\, d(\varphi, \psi) \qquad \text{on } (0, \mu]. \tag{91}$$

Proof. First, we prove exactly in the same way as in (85) that

$$\begin{aligned} \left| \frac{1}{\sqrt{\varphi(x) - \varphi(y)}} - \frac{1}{\sqrt{\psi(x) - \psi(y)}} \right| &\leq \frac{4\,(x^{3/2} - y^{3/2})}{3\,\beta^{3/2}\,(x-y)^{3/2}}\, d(\varphi, \psi) \\ &\leq \frac{C}{\sqrt{x-y}}\, d(\varphi, \psi). \end{aligned} \tag{92}$$

Besides, we have

$$\begin{aligned} n_2^1(\varphi)(x) - n_2^1(\psi)(x) = &\int_0^x [n_1(\varphi)(y) - n_1(\psi)(y)] \frac{g_\varphi(x,y)}{\sqrt{\varphi(x) - \varphi(y)}}\, dy \\ &+ \int_0^x n_1(\psi)(y)\, g_\varphi(x,y) \left[\frac{1}{\sqrt{\varphi(x) - \varphi(y)}} - \frac{1}{\sqrt{\psi(x) - \psi(y)}} \right] dy \\ &+ \int_0^x \frac{n_1(\psi)(y)}{\sqrt{\psi(x) - \psi(y)}} [g_\varphi(x,y) - g_\psi(x,y)]\, dy. \end{aligned} \tag{93}$$

From estimates (82) and (92), equation (93) can be estimated as follows:

$$\begin{aligned} \left|n_2^1(\varphi)(x) - n_2^1(\psi)(x)\right| \leq &\int_0^x \frac{k\, d(\varphi, \psi)}{4\sqrt{y}\sqrt{\beta(x-y)}}\, dy \\ &+ \int_0^x \frac{1}{\sqrt{\beta y}} \frac{C\, d(\varphi, \psi)}{\sqrt{x-y}}\, dy \\ &+ \int_0^x \frac{1}{\sqrt{\beta y}\sqrt{\beta(x-y)}} \left[\int_y^x \frac{C\, d(\varphi, \psi)}{\sqrt{h-y}}\, dh \right] dy. \end{aligned} \tag{94}$$

A straightforward computation of the right-hand side integrals leads to

$$\left|n_2^1(\varphi)(x) - n_2^1(\psi)(x)\right| \leq C\, d(\varphi, \psi), \qquad x \in (0, \mu]$$

and this ends the proof of Lemma 4.6. □

We proceed by induction and prove the following lemma:

LEMMA 4.7.

$$|n_2^p(\varphi)(x) - n_2^p(\psi)(x)| \le C\, h_p\left(\sqrt{\frac{x}{\beta}}\right) d(\varphi,\psi), \tag{95}$$

where

$$h_p(z) = \sum_{i=0}^{p-1} z^i \left(\prod_{j=1}^{i} I_j\right)$$

and

$$I_j = \int_0^1 \frac{t^{\frac{j-1}{2}}}{\sqrt{1-t}}\, dt.$$

Proof. We write

$$\begin{aligned} n_2^{p+1}(\varphi)(x) - n_2^{p+1}(\psi)(x) &= n_2^1(\varphi)(x) - n_2^1(\psi)(x) \\ &+ \int_0^x [n_2^p(\varphi)(y) - n_2^p(\psi)(y)] \frac{g_\varphi(x,y)}{\sqrt{\varphi(x)-\varphi(y)}}\, dy \\ &+ \int_0^x n_2^p(\psi)(y)\, g_\varphi(x,y) \left[\frac{1}{\sqrt{\varphi(x)-\varphi(y)}} - \frac{1}{\sqrt{\psi(x)-\psi(y)}}\right] dy \\ &+ \int_0^x \frac{n_2^p(\psi)(y)}{\sqrt{\psi(x)-\psi(y)}} [g_\varphi(x,y) - g_\psi(x,y)]\, dy. \end{aligned} \tag{96}$$

By using Lemma 4.6 and (92), we have

$$\begin{aligned} \left|n_2^{p+1}(\varphi)(x) - n_2^{p+1}(\psi)(x)\right| &\le C\, d(\varphi,\psi) \\ &+ \int_0^x \frac{|n_2^p(\varphi)(y) - n_2^p(\psi)(y)|}{\sqrt{\beta(x-y)}}\, dy \\ &+ C\|n_2^p(\psi)\|_{L^\infty} \sqrt{x}\, d(\varphi,\psi) \\ &+ C\|n_2^p(\psi)\|_{L^\infty} x\, d(\varphi,\psi). \end{aligned}$$

For μ small, we get from the preceding inequality

$$\left|n_2^{p+1}(\varphi)(x) - n_2^{p+1}(\psi)(x)\right| \le C_1\, d(\varphi,\psi) + \int_0^x \frac{|n_2^p(\varphi)(y) - n_2^p(\psi)(y)|}{\sqrt{\beta(x-y)}}\, dy \tag{97}$$

and the result follows easily by induction. □

End of the proof of Proposition 4.5. The inequality (95) passes to the limit and gives

$$|n_2(\varphi)(x) - n_2(\psi)(x)| \le C\, h\left(\sqrt{\frac{x}{\beta}}\right) d(\varphi,\psi), \tag{98}$$

where

$$h(z) = \sum_{i=0}^{+\infty} z^i \left(\prod_{j=1}^{i} I_j\right)$$

is an analytic function defined on $\mathbb{R}$. For μ small enough, we get

$$|n_2(\varphi) - n_2(\psi)| \leq \frac{k}{4\sqrt{x}} d(\varphi, \psi) \quad \forall x \in]0, \mu],$$

and this ends the proof. □

End of the proof of Proposition 4.2. By integrating twice the inequalities (82) and (83), we prove the strict contractivity of the map $\mathcal{S}$ for small μ which yields the existence and uniqueness of the solution on a small interval near zero. □

4.2. Local existence and uniqueness in the case $\beta = 0$. The proof will be analogous to the case $\beta > 0$ but the estimate (73) is not useful; thus we begin by proving the following lemma.

LEMMA 4.8. *There exists a positive μ such that for every solution φ of the system* (71) *with $\beta = 0$, the following estimate holds for every $x \in [0, \mu]$:*

$$\varphi(x) \geq \left(\frac{9}{8}\right)^{2/3} x^{4/3}. \tag{99}$$

Proof. Let φ be a solution of (71). Then the function $g_\varphi(x, 0)$ tends to 1 as x tends to zero. Thus, there exists a constant $\mu_0 > 0$ (depending on φ) such that

$$g_\varphi(x, 0) \geq \frac{1}{2} \quad \forall x \in [0, \mu_0]. \tag{100}$$

Since n_2 is positive, the above estimate gives

$$\frac{d^2\varphi}{dx^2} \geq \frac{1}{2\sqrt{\varphi}} \quad \text{on } [0, \mu_0]. \tag{101}$$

By multiplying this inequality by $\frac{d\varphi}{dx}$ and integrating, we get the estimate (99).

Now the only thing left to show is that μ_0 can be chosen independently of φ.

Let $\mu_1 \geq \mu_0$. For $x \in [\mu_0, \mu_1]$ the following estimate holds:

$$\begin{aligned} \log g_\varphi(x, 0) &= -\int_0^{\mu_0} \frac{dh}{\sqrt{\varphi(h)}} - \int_{\mu_0}^{x} \frac{dh}{\sqrt{\varphi(h)}} \\ &\geq -\int_0^{\mu_0} \frac{dh}{(\frac{9}{8})^{1/3} h^{2/3}} - \int_{\mu_0}^{x} \frac{dh}{\sqrt{\varphi(\mu_0)}} \\ &\geq -(\frac{8}{9})^{1/3} \left(2\,\mu_0^{1/3} + \frac{x}{\mu_0^{2/3}}\right). \end{aligned} \tag{102}$$

The estimate (99) will hold on $[0, \mu_1]$ if $g_\varphi(\mu_1, 0) \geq \frac{1}{2}$. Using the above estimate it is sufficient to take

$$\mu_1 = K\mu_0^{2/3} - 2\mu_0, \qquad K = \left(\frac{9}{8}\right)^{1/3} \log 2. \tag{103}$$

In the case $\mu_0 \geq (\frac{K}{3})^3$ then we take $\mu = \mu_0$. Otherwise, μ_1 defined by formula (103) satisfies

$$\mu_0 \leq \mu_1 \leq \left(\frac{K}{3}\right)^3.$$

Thus formula (103) permits us to build a bounded increasing sequence μ_n, with a limit equal to $(\frac{K}{3})^3$. In both cases we can choose $\mu = (\frac{K}{3})^3$, which ends the proof of the lemma. □

As in the case $\beta > 0$, the preceding lemma allows us to exhibit an equivalent of φ in the neighborhood of $x = 0$. Thus we obtain the analogue of formulae (75) and (76).

PROPOSITION 4.9. *There exist two positive constants μ and B such that the following inequalities hold on $[0, \mu]$ for every solution φ of (71) with $\beta = 0$:*

$$\frac{4}{9} D x^{-2/3} \left(1 - \frac{5}{2} B x^{\frac{1}{3}}\right) \leq \frac{d^2\varphi}{dx^2}(x) \leq \frac{4}{9} D x^{-2/3} \left(1 + \frac{5}{2} B x^{\frac{1}{3}}\right), \tag{104}$$

$$\frac{4}{3} D x^{1/3} \left(1 - \frac{5}{4} B x^{\frac{1}{3}}\right) \leq \frac{d\varphi}{dx}(x) \leq \frac{4}{3} D x^{1/3} \left(1 + \frac{5}{4} B x^{\frac{1}{3}}\right), \tag{105}$$

$$D x^{4/3}(1 - B x^{\frac{1}{3}}) \leq \varphi(x) \leq D x^{4/3}(1 + B x^{\frac{1}{3}}), \tag{106}$$

where

$$D = \left(\frac{9}{4}\right)^{2/3}.$$

Proof. Since φ is a convex function with $\varphi(0) = 0$, then for every $h \geq y \geq 0$ we have

$$\varphi(h) - \varphi(y) \geq \frac{h - y}{h} \varphi(h). \tag{107}$$

Thus using estimate (99), we get for $0 \leq y \leq x \leq \mu$

$$\exp\left(-\left(\frac{8}{9}\right)^{1/3} \int_y^x \frac{dh}{h^{1/6}\sqrt{h - y}}\right) \leq g_\varphi(x, y) \leq 1. \tag{108}$$

Now we apply this inequality for $y = 0$, and by using the inequality $\exp(-u) \geq 1 - u$, we get

$$\frac{1 - C x^{1/3}}{\sqrt{\varphi(x)}} \leq n_1(x) \leq \frac{1}{\sqrt{\varphi(x)}}. \tag{109}$$

Since (99) implies that $n_2 \in L^\infty_{4/3}[0, \mu]$, we have

$$n_2(x) \leq \frac{K}{x^{1/3}},$$

and therefore

$$\frac{1 - C x^{1/3}}{\sqrt{\varphi(x)}} \leq \frac{d^2\varphi}{dx^2}(x) \leq \frac{1}{\sqrt{\varphi(x)}} + \frac{K}{x^{1/3}}. \tag{110}$$

By using this estimate, we proceed as in [7] and prove the proposition. We build some sequences μ_n, B_n, C_n such that

$$\varphi(x) \geq C_n x^{4/3} (1 - B_n x^{1/3}) \quad \forall x \in [0, \mu_n]. \tag{111}$$

From the upper bound for $\frac{d^2\varphi}{dx^2}$ given by (110), we deduce

$$\frac{d^2\varphi}{dx^2}(x) \leq \frac{K}{x^{1/3}} + \frac{1}{\sqrt{C_n}\, x^{2/3}\sqrt{1 - B_n\, x^{1/3}}}. \tag{112}$$

Now we introduce a positive real number t such that

$$\frac{1}{\sqrt{1-u}} \leq 1 + u, \qquad \frac{1}{\sqrt{1+u}} \geq 1 - u, \qquad u \in [0, t]. \tag{113}$$

Thus for $x \in [0, \mu_n] \bigcap [0, (\frac{t}{B_n})^3]$, the estimate (112) turns to

$$\frac{d^2\varphi}{dx^2}(x) \leq \left(K + \frac{B_n}{\sqrt{C_n}}\right) x^{-1/3} + \frac{1}{\sqrt{C_n}\, x^{2/3}}. \tag{114}$$

We integrate this inequality twice, and by setting

$$C_n' = \frac{9}{4\sqrt{C_n}}, \quad B_n' = \frac{1}{C_n'}\left(\frac{9}{10}K + \frac{B_n}{\sqrt{C_n}}\right), \tag{115}$$

we get

$$\varphi(x) \leq C_n'\, x^{4/3}\,(1 + B_n'\, x^{1/3}), \qquad x \in [0, \mu_n] \bigcap [0, \left(\frac{t}{B_n}\right)^3]. \tag{116}$$

By using the above estimate and the first inequality of (110), we get

$$\frac{d^2\varphi}{dx^2}(x) \geq \frac{1 - Cx^{1/3}}{x^{2/3}\sqrt{C_n'}\sqrt{1 + B_n'\, x^{1/3}}}.$$

We integrate this inequality twice, using the estimate (113), and obtain

$$\varphi(x) \geq C_{n+1}\, x^{4/3}\,(1 - B_{n+1}\, x^{1/3}), \quad x \in [0, \mu_{n+1}], \tag{117}$$

where

$$\begin{aligned} C_{n+1} &= \frac{9}{4\sqrt{C_n'}}, \quad B_{n+1} = \frac{2}{5}(B_n' + C), \\ \mu_{n+1} &= \min\left[\mu_n, \left(\frac{t}{B_n}\right)^3, \left(\frac{t}{B_n'}\right)^3\right]. \end{aligned} \tag{118}$$

It is obvious to show that the sequences C_n and C_n' converge and that

$$\lim C_n = \lim C_n' = D = \left(\frac{9}{4}\right)^{2/3}. \tag{119}$$

By using (115) and (118), we can express B_{n+1} by means of B_n and show that B_n and B_n' converge, and that their limits are positive. Also, we can choose K and C in such a way that the limit B of B_n is the same as that of B_n'. This yields that μ_n converges to a positive limit μ. Thus the estimates (104)–(106) hold in $[0, \mu]$. □

Now we are able to define a set $\mathcal{E}'(\mu)$, where the solution lies, as shown in the following definition.

DEFINITION 4.10. *We denote by $\mathcal{E}'(\mu)$ the set of C^1 convex positive functions defined on $[0,\mu]$ such that*

$$\varphi(0) = 0, \quad \frac{d\varphi}{dx}(0) = 0, \tag{120}$$

$$\frac{1}{x^{2/3}}\left|\frac{d\varphi}{dx}(x) - \frac{4}{3}D\,x^{1/3}\right| \leq \frac{5}{3}DB \quad \text{on } (0,\mu).$$

On this set, we define the distance

$$d'(\varphi_1,\varphi_2) = \sup_{x\in(0,\mu]} \frac{1}{x^{2/3}}\left|\frac{d\varphi_1}{dx}(x) - \frac{d\varphi_2}{dx}(x)\right| \tag{121}$$

and then we define the map

$$\begin{aligned} \mathcal{S}:\ \mathcal{E}'(\mu) &\to \mathcal{E}'(\mu), \\ \varphi &\longmapsto \mathcal{S}(\varphi), \end{aligned} \tag{122}$$

such that

$$\begin{cases} \dfrac{d^2\mathcal{S}(\varphi)}{dx^2} = n(\varphi)(x), \\ n(x) = n_1(\varphi)(x) + n_2(\varphi)(x), \\ \mathcal{S}(\varphi)(0) = 0, \quad \dfrac{d\mathcal{S}(\varphi)}{dx}(0) = 0, \end{cases} \tag{123}$$

where $n_1(\varphi)$ and $n_2(\varphi)$ are defined by the formulae (57) *and* (58).

We claim that $\mathcal{S}$ maps $\mathcal{E}'(\mu)$ into itself because the constant B defined in the previous proposition is the limit of both sequences B_n and B'_n. Hence, one can do the same computations as in the proof of the previous proposition and prove the claim. Analogously to the previous section, we have the following lemma.

LEMMA 4.11. *The set $\mathcal{E}'(\mu)$ equipped with the distance d' is a complete metric space.*

Then we have the following proposition.

PROPOSITION 4.12. *There exist $\mu > 0$ and $k < 1$ such that for every function φ, ψ in $\mathcal{E}'(\mu)$, the following estimates hold on $(0,\mu]$:*

$$|n_1(\varphi)(x) - n_1(\psi)(x)| \leq \frac{k}{3x^{1/3}}\,d'(\varphi,\psi), \tag{124}$$

$$|n_2(\varphi)(x) - n_2(\psi)(x)| \leq \frac{k}{3x^{1/3}}\,d'(\varphi,\psi). \tag{125}$$

Proof. The proof of this proposition is analogous to that of Proposition 4.5. We begin by proving the following lemma. □

LEMMA 4.13. *There exists a constant $k_1 < 1$ such that for all φ and ψ in $\mathcal{E}'(\mu)$, the following estimate holds for $0 \leq y < x \leq \mu$:*

$$\left|\frac{1}{\sqrt{\varphi(x)-\varphi(y)}} - \frac{1}{\sqrt{\psi(x)-\psi(y)}}\right| \leq \frac{k_1}{3}\,\frac{x^{5/3}-y^{5/3}}{(x^{4/3}-y^{4/3})^{3/2}}\,d'(\varphi,\psi). \tag{126}$$

Proof. Since

$$\left|\frac{1}{\sqrt{\varphi(x)-\varphi(y)}}-\frac{1}{\sqrt{\psi(x)-\psi(y)}}\right| = \frac{\left|\int_y^x (\varphi'(h)-\psi'(h))\,dh\right|}{\sqrt{\varphi(x)-\varphi(y)}\sqrt{\psi(x)-\psi(y)}\left(\sqrt{\psi(x)-\psi(y)}+\sqrt{\varphi(x)-\varphi(y)}\right)},$$

we take μ small enough such that $(1-\frac{5}{4}B\mu^{1/3})^{3/2} \geq \frac{9}{10}$; therefore, (105) gives

$$\sqrt{\varphi(x)-\varphi(y)}\sqrt{\psi(x)-\psi(y)}\left(\sqrt{\psi(x)-\psi(y)}+\sqrt{\varphi(x)-\varphi(y)}\right) \geq 2\frac{9}{10}\left(\frac{4}{3}D\int_y^x h^{\frac{1}{3}}\,dh\right)^{3/2}.$$

Thus

$$\left|\frac{1}{\sqrt{\varphi(x)-\varphi(y)}}-\frac{1}{\sqrt{\psi(x)-\psi(y)}}\right| \leq d'(\varphi,\psi)\frac{\int_y^x h^{2/3}\,dh}{2\frac{9}{10}\left(\frac{4}{3}D\int_y^x h^{\frac{1}{3}}\,dh\right)^{3/2}}.$$

This proves the result with $k_1 = \frac{4}{9}$. □

By going back to the proof of Proposition 4.12, we prove (124):

$$n_1(\varphi)(x)-n_1(\psi)(x) = g_\varphi(x,0)\left[\frac{1}{\sqrt{\varphi(x)}}-\frac{1}{\sqrt{\psi(x)}}\right] + \frac{1}{\sqrt{\psi(x)}}\left[g_\varphi(x,0)-g_\psi(x,0)\right].$$

Therefore

$$|n_1(\varphi)(x)-n_1(\psi)(x)| \leq \frac{k_1}{3}x^{-1/3}\,d'(\varphi,\psi) + \frac{|g_\varphi(x,0)-g_\psi(x,0)|}{D^{3/2}\,x^{2/3}\sqrt{1-Bx^{1/3}}}.$$

In view of (126), the following estimate holds:

$$|g_\varphi(x,0)-g_\psi(x,0)| \leq \int_0^x \left|\frac{1}{\sqrt{\varphi(h)}}-\frac{1}{\sqrt{\psi(h)}}\right| dh \leq \frac{k_1}{2}\,x^{2/3}\,d'(\varphi,\psi).$$

Then

$$|n_1(\varphi)(x)-n_1(\psi)(x)| \leq \frac{k_1}{3}x^{-1/3}\,d'(\varphi,\psi) + Cd'(\varphi,\psi).$$

By taking $k_1 < k < 1$, we can choose μ small enough so that

$$|n_1(\varphi)(x)-n_1(\psi)(x)| \leq \frac{k}{3}x^{-1/3}\,d'(\varphi,\psi).$$

This ends the proof of (124). To prove (125), we proceed in the exact same manner as in the proof of (83). We consider the following sequences:

$$n_2^p(\varphi) = K_\varphi(n_2^{p-1}(\varphi)), \quad n_2^0(\varphi) = 0, \tag{127}$$

and

$$n_2^p(\psi) = K_\psi(n_2^{p-1}(\psi)), \quad n_2^0(\psi) = 0. \tag{128}$$

Therefore, we have the following lemma.

LEMMA 4.14. *There exist constants $C > 0$, $\mu > 0$ such that*

$$\left|n_2^1(\varphi) - n_2^1(\psi)\right| \le C\, d'(\varphi, \psi), \qquad \text{on } (0, \mu]. \tag{129}$$

Proof. We have

$$\begin{aligned} n_2^1(\varphi)(x) - n_2^1(\psi)(x) &= \int_0^x [n_1(\varphi)(y) - n_1(\psi)(y)] \frac{g_\varphi(x,y)}{\sqrt{\varphi(x) - \varphi(y)}}\, dy \\ &\quad + \int_0^x n_1(\psi)(y)\, g_\varphi(x,y) \left[\frac{1}{\sqrt{\varphi(x) - \varphi(y)}} - \frac{1}{\sqrt{\psi(x) - \psi(y)}}\right] dy \\ &\quad + \int_0^x \frac{n_1(\psi)(y)}{\sqrt{\psi(x) - \psi(y)}} [g_\varphi(x,y) - g_\psi(x,y)]\, dy. \end{aligned} \tag{130}$$

From estimates (124) and (126), the right-hand side of equation (130) can be estimated as follows:

$$\begin{aligned} \left|n_2^1(\varphi)(x) - n_2^1(\psi)(x)\right| &\le \int_0^x \frac{k\, d'(\varphi,\psi)}{3y^{1/3}\sqrt{C(x^{4/3} - y^{4/3})}}\, dy \\ &\quad + \int_0^x \frac{1}{\sqrt{Cy^{4/3}}} \frac{k_1}{3} \frac{x^{5/3} - y^{5/3}}{(x^{4/3} - y^{4/3})^{3/2}}\, d'(\varphi,\psi)\, dy \\ &\quad + \int_0^x \frac{1}{\sqrt{Cy^{4/3}}\sqrt{C(x^{4/3} - y^{4/3})}} \\ &\qquad \left[\int_y^x C\, d'(\varphi,\psi) \frac{k_1}{3} \frac{h^{5/3} - y^{5/3}}{(y^{4/3} - y^{4/3})^{3/2}} dh\right] dy. \end{aligned} \tag{131}$$

By the change of the variable $y = xt$, we prove that the first two integrals of the right-hand side are constant. To compute the third integral, we set

$$A = \int_0^x \frac{1}{\sqrt{Cy^{4/3}}\sqrt{C(x^{4/3} - y^{4/3})}} \left[\int_y^x C \frac{k_1}{3} \frac{h^{5/3} - y^{5/3}}{(h^{4/3} - y^{4/3})^{3/2}}\, dh\right] dy.$$

Then, through Fubini's formula, A satisfies

$$A = C \int_0^x \frac{1}{\sqrt{x^{4/3} - h^{4/3}}} \int_0^h \frac{h^{5/3} - y^{5/3}}{\sqrt{y^{4/3}}\,(h^{4/3} - y^{4/3})^{3/2}}\, dy\, dh.$$

By using the change of variable $y = ht$ for fixed h , we obtain

$$A = CI \int_0^x \frac{dh}{\sqrt{x^{4/3} - h^{4/3}}},$$

where

$$I = \int_0^1 \frac{1 - t^{5/3}}{t^{2/3}(1 - t^{4/3})^{3/2}}\, dt < \infty.$$

Again by the change of variable $h = tx$, we obtain

$$A = C\, x^{1/3},$$

and since the third term of (131) is equal to $A\, d'(\varphi, \psi)$, then

$$\left| n_2^1(\varphi)(x) - n_2^1(\psi)(x) \right| \leq C\, d'(\varphi, \psi), \qquad x \in (0, \mu]$$

and this ends the proof of Lemma 4.14. □

We use this lemma and proceed by induction to prove the following lemma.

LEMMA 4.15.

$$|n_2^p(\varphi)(x) - n_2^p(\psi)(x)| \leq C\, h_p((Cx)^{1/3})\, d'(\varphi, \psi), \tag{132}$$

where

$$h_p(z) = \sum_{i=0}^{p-1} z^i \left(\prod_{j=1}^{i} I_j \right)$$

and

$$I_j = \int_0^1 \frac{t^{\frac{j-1}{3}}}{\sqrt{1 - t^{4/3}}}\, dt.$$

Proof. We write

$$\begin{aligned} n_2^{p+1}(\varphi)(x) - n_2^{p+1}(\psi)(x) &= n_2^1(\varphi)(x) - n_2^1(\psi)(x) \\ &+ \int_0^x [n_2^p(\varphi)(y) - n_2^p(\psi)(y)] \frac{g_\varphi(x, y)}{\sqrt{\varphi(x) - \varphi(y)}}\, dy \\ &+ \int_0^x n_2^p(\psi)(y)\, g_\varphi(x, y) \left[\frac{1}{\sqrt{\varphi(x) - \varphi(y)}} - \frac{1}{\sqrt{\psi(x) - \psi(y)}} \right] dy \\ &+ \int_0^x \frac{n_2^p(\psi)(y)}{\sqrt{\psi(x) - \psi(y)}} [g_\varphi(x, y) - g_\psi(x, y)]\, dy \end{aligned} \tag{133}$$

and by using Lemma 4.14 and inequalities (87) and (126), we deduce that

$$\begin{aligned} \left| n_2^{p+1}(\varphi)(x) - n_2^{p+1}(\psi)(x) \right| &\leq C\, d'(\varphi, \psi) \\ &+ \int_0^x \frac{|n_2^p(\varphi)(y) - n_2^p(\psi)(y)|}{\sqrt{C(x^{4/3} - y^{4/3})}}\, dy \\ &+ C x^{1/3}\, d'(\varphi, \psi) \\ &+ C x^{2/3}\, d'(\varphi, \psi). \end{aligned}$$

For μ small, we get from the preceding inequality that

$$\left| n_2^{p+1}(\varphi)(x) - n_2^{p+1}(\psi)(x) \right| \leq C_1\, d'(\varphi, \psi) + \int_0^x \frac{|n_2^p(\varphi)(y) - n_2^p(\psi)(y)|}{\sqrt{C(x^{4/3} - y^{4/3})}}\, dy, \tag{134}$$

and the result follows easily by induction. □

End of the proof of Proposition 4.12. The inequality (132) of Lemma 4.15 passes to the limit and gives

$$|n_2(\varphi)(x) - n_2(\psi)(x)| \leq C\, h((Cx)^{1/3})\, d'(\varphi, \psi), \tag{135}$$

where

$$h(z) = \sum_{i=0}^{+\infty} z^i \left(\prod_{j=1}^{i} I_j \right)$$

is an analytic function defined on $\mathbb{R}$. For μ small enough we get

$$|n_2(\varphi)(x) - n_2(\psi)(x)| \leq \frac{k}{3x^{1/3}}, d'(\varphi, \psi) \quad \forall x \in (0, \mu]$$

and this ends the proof. □

End of the proof of Proposition 4.2. Like in the case $\beta > 0$ treated in the previous section we integrate twice the inequalities (124) and (125) and prove the strict contractivity of the map $\mathcal{S}$ for small μ. This gives the existence and the uniqueness of the solution on a small interval near zero. □

4.3. Global existence and uniqueness. In this section, we show the existence and uniqueness of the solution of the system (71) on an interval $[\mu, A]$, where A is an arbitrary constant. By denoting $\widetilde{\varphi}, \widetilde{n}$ as the quantities defined in Proposition 4.2, we can rewrite the system on $[\mu, A]$:

$$\frac{d^2\varphi}{dx^2} = n(x), \qquad x \in [\mu, A], \tag{136}$$

$$n(x) = n_1(x) + n_2(x), \tag{137}$$

$$n_1(x) = \frac{1}{\sqrt{\varphi(x)}}\, g_{\widetilde{\varphi}}(\mu, 0) \exp\left(- \int_{\mu}^{x} \frac{dz}{\sqrt{\varphi(z)}} \right), \tag{138}$$

$$\begin{aligned} n_2(x) &- \int_0^{\mu} \frac{\widetilde{n}(y)}{\sqrt{\varphi(x) - \widetilde{\varphi}(y)}} g_{\widetilde{\varphi}}(\mu, y) \exp\left(- \int_{\mu}^{x} \frac{dz}{\sqrt{\varphi(z) - \widetilde{\varphi}(y)}} \right) dy \\ &- \int_{\mu}^{x} \frac{g_{\varphi}(x, y)}{\sqrt{\varphi(x) - \varphi(y)}} \left[n_1(y) + n_2(y) \right] dy \\ &= 0, \end{aligned} \tag{139}$$

$$\varphi(\mu) = \widetilde{\varphi}(\mu), \qquad \frac{d\varphi}{dx}(\mu) = \frac{d\widetilde{\varphi}}{dx}(\mu), \tag{140}$$

$$g_{\varphi}(x, y) = \exp\left(- \int_{y}^{x} \frac{dz}{\sqrt{\varphi(z) - \varphi(y)}} \right), \tag{141}$$

$$g_{\widetilde{\varphi}}(x,y) = \exp\Big(- \int_y^x \frac{dz}{\sqrt{\widetilde{\varphi}(z) - \widetilde{\varphi}(y)}} \Big). \tag{142}$$

We apply a fixed-point procedure associated with this new formulation. We set

$$\mathcal{F} = \left\{ \varphi \in C^1[\mu, A] \text{ convex such that } \quad \varphi(\mu) = \widetilde{\varphi}(\mu), \quad \frac{d\varphi}{dx}(\mu) = \frac{d\widetilde{\varphi}}{dx}(\mu) \right\}.$$

We define the map

$$n_1(\varphi)(x) = \frac{1}{\sqrt{\varphi(x)}}\, g_{\widetilde{\varphi}}(\mu, 0) \exp\Big(- \int_\mu^x \frac{dz}{\sqrt{\varphi(z)}} \Big)$$

and the map $n_2(\varphi)$ by the integral equation

$$\begin{aligned} n_2(\varphi)(x) &- \int_0^\mu \frac{\widetilde{n}(y)}{\sqrt{\varphi(x) - \widetilde{\varphi}(y)}} g_{\widetilde{\varphi}}(\mu, y) \exp\Big(- \int_\mu^x \frac{dz}{\sqrt{\varphi(z) - \widetilde{\varphi}(y)}} \Big) dy \\ &- \int_\mu^x \frac{g_\varphi(x,y)}{\sqrt{\varphi(x) - \varphi(y)}} \left[n_1(\varphi)(y) + n_2(\varphi)(y) \right] dy. \\ &= 0. \end{aligned} \tag{143}$$

Finally, we define the map $\mathcal{S}$

$$\begin{aligned} \mathcal{S} :\ & \mathcal{F} \to \mathcal{F}, \\ & \varphi \longmapsto \mathcal{S}(\varphi) \end{aligned} \tag{144}$$

such that

$$\begin{cases} \dfrac{d^2 \mathcal{S}(\varphi)}{dx^2} = n(\varphi)(x), \\ n(\varphi)(x) = n_1(\varphi)(x) + n_2(\varphi)(x), \\ \mathcal{S}(\varphi)(\mu) = \widetilde{\varphi}(\mu), \qquad \dfrac{d\mathcal{S}(\varphi)}{dx}(\mu) = \dfrac{d\widetilde{\varphi}}{dx}(\mu) > 0. \end{cases} \tag{145}$$

The proof of existence and uniqueness of the solution is based on the following proposition.

PROPOSITION 4.16. *There exists a constant k such that for every pair of functions φ, ψ in $\mathcal{F}$ and every integer p that satisfies*

$$|\varphi'(x) - \psi'(x)| \le D_1 (x - \mu)^p \qquad \forall x \in [\mu, A],$$

where D_1 is an arbitrary constant, we have

$$|\mathcal{S}(\varphi)'(x) - \mathcal{S}(\psi)'(x)| \le \frac{k D_1}{p+1} (x - \mu)^{p+1} \qquad \forall x \in [\mu, A].$$

The existence and uniqueness of the solution is an immediate consequence of the above proposition.

COROLLARY 4.17. *There exists an integer p such that the map $\mathcal{S}^p$ is a contraction on $\mathcal{F}$. The considered distance on $\mathcal{F}$ is*

$$d(\varphi, \psi) = \sup_{x \in [\mu, A]} |\varphi'(x) - \psi'(x)|.$$

Proof. Apply the preceding proposition with $p = 0$, then

$$|\mathcal{S}(\varphi)'(x) - \mathcal{S}(\psi)'(x)| \leq k\,(x-\mu)\,d(\varphi,\psi) \qquad \forall x \in [\mu, A].$$

An iteration of this inequality using the preceding proposition gives

$$|\mathcal{S}^p(\varphi)'(x) - \mathcal{S}^p(\psi)'(x)| \leq \frac{k^p\,(x-\mu)^p}{p!}\,d(\varphi,\psi) \qquad \forall x \in [\mu, A].$$

This gives

$$d(\mathcal{S}^p(\varphi), \mathcal{S}^p(\psi)) \leq \frac{k^p\,(A-\mu)^p}{p!}\,d(\varphi,\psi),$$

and we only have to choose p large enough. □

The remainder of this paragraph will consist in proving Proposition 4.16. From now on we will consider a pair φ, ψ of functions in $\mathcal{F}$ that satisfy

$$|\varphi'(x) - \psi'(x)| \leq D_1\,(x-\mu)^p \qquad \forall x \in [\mu, A]. \tag{146}$$

LEMMA 4.18. *There exists a constant k_1 independent of φ, ψ, and p such that*

$$|n_1(\varphi)(x) - n_1(\psi)(x)| \leq k_1\,D_1\,(x-\mu)^p.$$

Proof. The difference of the densities associated, respectively, with φ and ψ reads

$$\begin{aligned} n_1(\varphi)(x) - n_1(\psi)(x) &= \frac{g_{\widetilde{\varphi}}(\mu,0)}{\sqrt{\varphi(x)}}\left[\exp\Big(-\int_\mu^x \frac{dz}{\sqrt{\varphi(z)}}\Big) - \exp\Big(-\int_\mu^x \frac{dz}{\sqrt{\psi(z)}}\Big)\right] \\ &\quad + \left(\frac{1}{\sqrt{\varphi(x)}} - \frac{1}{\sqrt{\psi(x)}}\right) g_{\widetilde{\varphi}}(\mu,0)\,\exp\Big(-\int_\mu^x \frac{dz}{\sqrt{\psi(z)}}\Big). \end{aligned} \tag{147}$$

The second term of the right-hand side can be bounded by

$$\left|\frac{1}{\sqrt{\varphi(x)}} - \frac{1}{\sqrt{\psi(x)}}\right| \leq \frac{|\varphi(x) - \psi(x)|}{\sqrt{\varphi}\,\sqrt{\psi}\,(\sqrt{\varphi} + \sqrt{\psi})}.$$

By the convexity of φ and ψ, we have $\varphi, \psi \geq \widetilde{\varphi}(\mu) > 0$. Then

$$\left|\frac{1}{\sqrt{\varphi(x)}} - \frac{1}{\sqrt{\psi(x)}}\right| \leq \frac{1}{2(\widetilde{\varphi}(\mu))^{3/2}} \int_\mu^x |\varphi'(s) - \psi'(s)|\,ds \leq k_1'\,D_1\,(x-\mu)^p,$$

where the constant k_1' only depends on μ, $\widetilde{\varphi}(\mu)$ and A.

The first term of (147) can be estimated as follows:

$$\begin{aligned} &\frac{g_{\widetilde{\varphi}}(\mu,0)}{\sqrt{\varphi(x)}}\left[\exp\Big(-\int_\mu^x \frac{dz}{\sqrt{\varphi(z)}}\Big) - \exp\Big(-\int_\mu^x \frac{dz}{\sqrt{\psi(z)}}\Big)\right]. \\ &\leq \frac{1}{\sqrt{\widetilde{\varphi}(\mu)}} \int_\mu^x \left|\frac{1}{\sqrt{\varphi(h)}} - \frac{1}{\sqrt{\psi(h)}}\right| dh \\ &\leq k_1''\,D_1\,(x-\mu)^p \end{aligned} \tag{148}$$

This inequality, combined with the preceding one, ends the proof of the lemma. □

We will prove an analogous lemma for the density n_2. Like previous sections, we will build a sequence n_2^k, prove some estimates on n_2^k, and then pass to the limit.

We set

$$n_2^0(\varphi) = 0, \tag{149}$$

and for all k,

$$\begin{aligned} n_2^{k+1}(\varphi)(x) = &\int_0^\mu \frac{\widetilde{n}(y)}{\sqrt{\varphi(x)-\widetilde{\varphi}(y)}} g_{\widetilde{\varphi}}(\mu,y)\, \exp\Big(-\int_\mu^x \frac{dz}{\sqrt{\varphi(z)-\widetilde{\varphi}(y)}}\Big)dy \\ &+\int_\mu^x \frac{g_\varphi(x,y)}{\sqrt{\varphi(x)-\varphi(y)}}\left[n_1(\varphi)(y)+n_2^k(\varphi)(y)\right] dy. \end{aligned} \tag{150}$$

Notice that

$$n_2^{k+1}(\varphi)(x) = n_2^1(\varphi)(x) + \int_\mu^x \frac{g_\varphi(x,y)\, n_2^k(\varphi)(y)}{\sqrt{\varphi(x)-\varphi(y)}}\, dy. \tag{151}$$

We begin with the following lemma.

LEMMA 4.19. *There exists a constant k_2^1 such that for every φ, ψ that satisfies* (146)*, the following estimate holds on* $[\mu, A]$:

$$|n_2^1(\varphi)(x) - n_2^1(\psi)(x)| \le k_2^1\, D_1\, (x-\mu)^p. \tag{152}$$

Proof. We write

$$n_2^1(\varphi)(x) - n_2^1(\psi)(x) = I + II + III + IV + V, \tag{153}$$

where

$$\begin{aligned} I = &\int_0^\mu \widetilde{n}(y) g_{\widetilde{\varphi}}(\mu,y)\, \exp\left(-\int_\mu^x \frac{dz}{\sqrt{\varphi(z)-\widetilde{\varphi}(y)}}\right) \\ &\times \left[\frac{1}{\sqrt{\varphi(x)-\widetilde{\varphi}(y)}} - \frac{1}{\sqrt{\psi(x)-\widetilde{\varphi}(y)}}\right] dy, \end{aligned} \tag{154}$$

$$\begin{aligned} II = &\int_0^\mu \frac{\widetilde{n}(y)\, g_{\widetilde{\varphi}}(\mu,y)}{\sqrt{\psi(x)-\widetilde{\varphi}(y)}} \left[\exp\left(-\int_\mu^x \frac{dz}{\sqrt{\varphi(z)-\widetilde{\varphi}(y)}}\right)\right. \\ &\left. - \exp\left(-\int_\mu^x \frac{dz}{\sqrt{\psi(z)-\widetilde{\varphi}(y)}}\right)\right] dy, \end{aligned} \tag{155}$$

$$III = \int_\mu^x \frac{n_1(\varphi)(y) - n_1(\psi)(y)}{\sqrt{\varphi(x)-\varphi(y)}}\, g_\varphi(x,y)\, dy, \tag{156}$$

$$IV = \int_\mu^x n_1(\psi)(y) g_\psi(x,y) \left[\frac{1}{\sqrt{\varphi(x)-\varphi(y)}} - \frac{1}{\sqrt{\psi(x)-\psi(y)}}\right] dy, \tag{157}$$

$$V = \int_\mu^x \frac{n_1(\psi)(y)}{\sqrt{\psi(x)-\psi(y)}} \left[g_\varphi(x,y) - g_\psi(x,y)\right] dy. \tag{158}$$

Now we estimate the five terms one by one. First, since from the previous section $\widetilde{n}(y)y^{2/3}$ is a bounded function in the neighborhood of zero, then the term I can be estimated as follows:

$$|I| \leq \int_0^{\mu} Ky^{-2/3} \left| \frac{1}{\sqrt{\varphi(x) - \widetilde{\varphi}(y)}} - \frac{1}{\sqrt{\psi(x) - \widetilde{\varphi}(y)}} \right| dy.$$

But since

$$\left| \frac{1}{\sqrt{\varphi(x) - \widetilde{\varphi}(y)}} - \frac{1}{\sqrt{\psi(x) - \widetilde{\varphi}(y)}} \right| = \frac{|\varphi(x) - \psi(x)|}{\sqrt{\varphi(x) - \widetilde{\varphi}(y)}\sqrt{\psi(x) - \widetilde{\varphi}(y)}(\sqrt{\varphi(x) - \widetilde{\varphi}(y)} + \sqrt{\psi(x) - \widetilde{\varphi}(y)})}$$

and

$$\varphi(x) - \widetilde{\varphi}(y) \geq \frac{d\widetilde{\varphi}}{dx}(\mu)(x - \mu) + \frac{d\widetilde{\varphi}}{dx}(y)(\mu - y) \geq C(x - y),$$

then

$$|I| \leq |\varphi(x) - \psi(x)| \int_0^{\mu} \frac{K' \, dy}{y^{2/3} \, (x - y)^{3/2}}.$$

Therefore,

$$|I| \leq C \frac{|\varphi(x) - \psi(x)|}{\sqrt{x - \mu}} \leq C \, D_1 \, (x - \mu)^p. \tag{159}$$

Analogously,

$$\begin{aligned} |II| &\leq \int_0^{\mu} \frac{C}{y^{2/3}\sqrt{x - y}} \int_{\mu}^{x} \left| \frac{1}{\sqrt{\varphi(z) - \widetilde{\varphi}(y)}} - \frac{1}{\sqrt{\psi(z) - \widetilde{\varphi}(y)}} \right| dz \, dy \\ &\leq C \int_0^{\mu} \frac{dy}{y^{2/3}\sqrt{x - y}} \int_{\mu}^{x} \frac{|\varphi(z) - \psi(z)| \, dz}{(z - y)^{3/2}} \\ &\leq C \int_{\mu}^{x} dz \, |\varphi(z) - \psi(z)| \int_0^{\mu} \frac{dy}{y^{2/3}\sqrt{x - y}\,(z - y)^{3/2}}. \end{aligned}$$

Therefore,

$$|II| \leq C \int_{\mu}^{x} \frac{|\varphi(z) - \psi(z)|}{\sqrt{z - \mu}\sqrt{x - \mu}} \, dz \leq C \, D_1 \, (x - \mu)^p. \tag{160}$$

For III, the estimate comes directly from the one on n_1,

$$|III| \leq \int_{\mu}^{x} \frac{C|n_1(\varphi)(y) - n_1(\psi)(y)|}{\sqrt{x - y}} \, dy.$$

We deduce easily that

$$|III| \leq C \, D_1 \, (x - \mu)^p. \tag{161}$$

For IV, we have, since n_1 is bounded in $L^\infty(\mu, A)$,

$$|IV| \leq C \int_\mu^x \left| \frac{1}{\sqrt{\varphi(x)-\varphi(y)}} - \frac{1}{\sqrt{\psi(x)-\psi(y)}} \right| dy.$$

But since

$$\left| \frac{1}{\sqrt{\varphi(x)-\varphi(y)}} - \frac{1}{\sqrt{\psi(x)-\psi(y)}} \right| \leq C \frac{\int_y^x |\varphi'(h)-\psi'(h)|\, dh}{(x-y)^{3/2}},$$

we use Fubini's theorem and obtain

$$|IV| \leq \int_\mu^x dh\, |\varphi'(h)-\psi'(h)| \int_\mu^h \frac{dy}{(x-y)^{3/2}}.$$

This leads to

$$|IV| \leq \int_\mu^x \frac{|\varphi'(h)-\psi'(h)|}{\sqrt{x-y}}\, dh \leq C\, D_1\, (x-\mu)^p. \tag{162}$$

For the last term, we obtain

$$\begin{aligned}
|V| &\leq \int_\mu^x \frac{C}{\sqrt{x-y}} \int_y^x \left| \frac{1}{\sqrt{\varphi(h)-\varphi(y)}} - \frac{1}{\sqrt{\psi(h)-\psi(y)}} \right| dh\, dy \\
&\leq \int_\mu^x \int_y^x \int_y^h \frac{C\, |\varphi'(t)-\psi'(t)|}{(h-y)^{3/2}\sqrt{x-y}}\, dt\, dh\, dy \\
&\leq \int_\mu^x \int_y^x \int_t^x \frac{C\, |\varphi'(t)-\psi'(t)|}{(h-y)^{3/2}\sqrt{x-y}}\, dh\, dt\, dy \\
&\leq \int_\mu^x \int_y^x \frac{C\, |\varphi'(t)-\psi'(t)|}{\sqrt{t-y}\sqrt{x-y}}\, dt\, dy \\
&\leq \int_\mu^x dt\, |\varphi'(t)-\psi'(t)| \int_\mu^t \frac{C}{\sqrt{t-y}\sqrt{x-y}}\, dy \\
&\leq \int_\mu^x C\, |\varphi'(t)-\psi'(t)| \frac{\sqrt{t-\mu}}{\sqrt{x-t}}\, dt.
\end{aligned}$$

Thus, we have

$$|V| \leq C \int_\mu^x \frac{|\varphi'(t)-\psi'(t)|}{\sqrt{x-t}}\, dt \leq C\, D_1\, (x-\mu)^p. \tag{163}$$

The result follows easily from (159)–(163). □

By using the sequences n_2^k, we prove the following proposition.

PROPOSITION 4.20. *There exists a constant k_2 such that for every φ, ψ in $\mathcal{F}$ that satisfies* (146), *the following estimate holds on* $[\mu, A]$:

$$|n_2(\varphi)(x) - n_2(\psi)(x)| \leq k_2\, D\, (x-\mu)^p.$$

Proof. We will prove this estimate for the difference $|n_2^k(\varphi)(x) - n_2^k(\psi)(x)|$ and then pass to the limit. From (151), we deduce that

$$
\begin{aligned}
n_2^{k+1}(\varphi)(x) - n_2^{k+1}(\psi)(x) &= n_2^1(\varphi)(x) - n_2^1(\psi)(x) \\
&+ \int_\mu^x \frac{n_2^k(\varphi)(y) - n_2^k(\psi)(y)}{\sqrt{\varphi(x) - \varphi(y)}}\, g_\varphi(x,y)dy \\
&+ \int_\mu^x n_2^k(\psi)(y)\, g_\varphi(x,y) \left[\frac{1}{\sqrt{\varphi(x) - \varphi(y)}} - \frac{1}{\sqrt{\psi(x) - \psi(y)}}\right] dy \\
&+ \int_\mu^x \frac{n_2^k(\psi)(y)}{\sqrt{\psi(x) - \psi(y)}}\, [g_\varphi(x,y) - g_\psi(x,y)]\, dy.
\end{aligned}
$$

Since n_2^k can be bounded independently of φ and ψ on $[\mu, A]$, then the last two terms of the above equation can be treated exactly as the terms IV and V of the proof of Lemma 4.19. (See (157) and (158).) By using the result of Lemma 4.18, we can write

$$
\begin{aligned}
&\left|n_2^{k+1}(\varphi)(x) - n_2^{k+1}(\psi)(x)\right| \\
&\quad \le C_1 \left[D_1\,(x-\mu)^p + \int_\mu^x \frac{|n_2^k(\varphi)(y) - n_2^k(\psi)(y)|}{\sqrt{x-y}}\, dy\right].
\end{aligned} \tag{164}
$$

An immediate iteration gives

$$
\left|n_2^k(\varphi)(x) - n_2^k(\psi)(x)\right| \le C_1\, D_1\, (x-\mu)^p\, h_{p,k}\,(C_1\sqrt{x-\mu}), \tag{165}
$$

where

$$
h_{p,k}(z) = \sum_{j=0}^{k-1} z^j \left(\prod_{i=1}^{j} I_{2p+i}\right) \tag{166}
$$

with

$$
I_m = \int_0^1 \frac{t^{m/2}}{\sqrt{1-t}}\, dt, \quad m \in \mathbb{N}. \tag{167}
$$

Since $h_{p,k}\,(C_1\sqrt{x-\mu}) \le h_{0.k}\,(C_1\sqrt{A-\mu})$, and since this sequence is convergent as k tends to infinity, we deduce that

$$
|n_2(\varphi)(x) - n_2(\psi)(x)| \le k_2\, D_1\, (x-\mu)^p. \qquad \square
$$

5. The boundary-value problem and the Child–Langmuir current. In the previous sections we have shown that the Cauchy problem (71) introduced in §4 has a unique solution. After performing the inverse scaling of (54), we obtained that the problem

$$
\frac{d^2\varphi}{dx^2} = n(x), \qquad n(x) = n_1(x) + n_2(x), \tag{168}
$$

$$
n_1(x) = \frac{j}{\sqrt{\varphi(x)}}\, g_\varphi(x,0), \tag{169}
$$

$$
\begin{aligned}
n_2(x) &- \int_0^x \frac{g_\varphi(x,y)}{\tau\sqrt{\varphi(x) - \varphi(y)}} n_2(y)\, dy, \\
&= \int_0^x \frac{g_\varphi(x,y)}{\tau\sqrt{\varphi(x) - \varphi(y)}} n_1(y)\, dy,
\end{aligned} \tag{170}
$$

$$\varphi(0) = 0, \qquad \frac{d\varphi}{dx}(0) = \beta_1, \tag{171}$$

$$g_\varphi(x,y) = \exp\left(-\frac{1}{\tau} \int_y^x \frac{dz}{\sqrt{\varphi(z) - \varphi(y)}} \right) \tag{172}$$

has a unique solution for every $\tau > 0$, $j \geq 0$, and $\beta_1 \geq 0$. This solution can be written as

$$\varphi(x) = \tau^4 j^2 \, \widetilde{\varphi}\left(\frac{x}{\tau^3 j} \right), \tag{173}$$

where $\widetilde{\varphi}$ is the solution of (71) with $\beta = \frac{\beta_1}{\tau j}$. Therefore φ is a solution of the boundary-value problem (55)–(60) if and only if $\varphi(1) = 1$ or, equivalently, if and only if the corresponding $\widetilde{\varphi}$ satisfies

$$\widetilde{\varphi}\left(\frac{1}{\tau^3 j} \right) = \frac{1}{\tau^4 j^2}. \tag{174}$$

This constraint will give the limitation of the current. Let us start with the following proposition.

PROPOSITION 5.1. *Let φ be the solution of the problem* (168)–(172); *then we have*

$$\int_0^x n(y) g_\varphi(x,y)\, dy = \tau j \,(1 - g_\varphi(x,0)) \tag{175}$$

and

$$\begin{aligned} \varphi'^2(x) = \beta_1^2 + \frac{4j}{\tau} x + 4j \sqrt{\varphi(x)}\, g_\varphi(x,0) \\ + \frac{4}{\tau} \int_0^x n(y) g_\varphi(x,y) \sqrt{\varphi(x) - \varphi(y)}\, dy. \end{aligned} \tag{176}$$

Proof. (46) shows that

$$\begin{aligned} \int_0^x n(y) g_\varphi(x,y)\, dy &= \tau \int_0^x \bar{j}_2(x,y) dy \\ &= \tau j_2(x); \end{aligned}$$

then with (45) and (39) we have

$$\int_0^x n(y) g_\varphi(x,y)\, dy = \tau j \,(1 - g_\varphi(x,0)),$$

which shows (175).

(176) can be shown by multiplying the Poisson equation by $\frac{d\varphi}{dx}$ and integrating it between 0 and x. This leads to

$$\begin{aligned} \varphi'^2(x) - \varphi'^2(0) = 2 \int_0^x \frac{j}{\sqrt{\varphi(y)}} g_\varphi(y,0) \varphi'(y)\, dy \\ + 2 \int_0^x \varphi'(y) \int_0^y \frac{n(z) g_\varphi(y,z)}{\tau \sqrt{\varphi(y) - \varphi(z)}} dz\, dy. \end{aligned} \tag{177}$$

The first integral can be computed by an integration by parts

$$\int_0^x \frac{j}{\sqrt{\varphi(y)}}\, g_\varphi(y,0)\varphi'(y)\,dy = 2jg_\varphi(x,0)\sqrt{\varphi(x)} + 2\frac{j}{\tau}\int_0^x g_\varphi(y,0)dy. \tag{178}$$

For the second integral, we use Fubini's theorem

$$\int_0^x \varphi'(y)\int_0^y \frac{n(z)g_\varphi(y,z)}{\tau\sqrt{\varphi(y)-\varphi(z)}}dz\,dy = \int_0^x n(z)\int_z^x \frac{\varphi'(y))g_\varphi(y,z)}{\tau\sqrt{\varphi(y)-\varphi(z)}}dy\,dz.$$

An integration by parts for fixed z gives

$$\int_0^x n(z)\int_z^x \frac{\varphi'(y)g_\varphi(y,z)}{\tau\sqrt{\varphi(y)-\varphi(z)}}dy\,dz = 2\int_0^x n(z)\frac{g_\varphi(x,z)}{\tau}\sqrt{\varphi(x)-\varphi(z)}dz + 2\int_0^x n(z)\int_z^x \frac{g_\varphi(y,z)}{\tau}dy\,dz.$$

By using Fubini's theorem backward for the last integral and applying (175), we obtain

$$\int_0^x \varphi'(y)\int_0^y \frac{n(z)g_\varphi(y,z)}{\tau\sqrt{\varphi(y)-\varphi(z)}}dz\,dy = 2\int_0^x n(z)\frac{g_\varphi(x,z)}{\tau}\sqrt{\varphi(x)-\varphi(z)}dz + 2\frac{j}{\tau}x - 2\frac{j}{\tau}\int_0^x g_\varphi(y,0)dy. \tag{179}$$

We sum this equation and equation (177) to finally obtain (176). □

Remark. (176) can also be derived from the Boltzmann equation (26) by multiplying by v and integrating. This can be viewed as a sort of energy identity.

This proposition allows us to prove Theorem 2.2.

Proof of Theorem 2.2. First, we deduce from (176) that

$$\varphi'^2(x) \geq \frac{4j}{\tau}x. \tag{180}$$

We take the square root of this inequality and integrate it. Hence the condition $\varphi(1) = 1$ leads to

$$j \leq \frac{9}{16}\tau. \tag{181}$$

Besides, by using the expression (170) of n_2, we have

$$n_2(x) = \int_0^x \frac{n(y)}{\tau\sqrt{\varphi(x)-\varphi(y)}}g_\varphi(x,y)\,dy > \frac{1}{\tau\sqrt{\varphi(x)}}\int_0^x n(y)g_\varphi(x,y)dy.$$

Hence, by using (175) we obtain

$$n_2(x) > \frac{j}{\sqrt{\varphi(x)}}(1-g_\varphi(x,0)) = \frac{j}{\sqrt{\varphi(x)}} - n_1(x).$$

Then (168) leads to

$$\frac{d^2\varphi}{dx^2} > \frac{j}{\sqrt{\varphi}}, \tag{182}$$

which by integration gives

$$j < 4/9 \tag{183}$$

and ends the proof. □

As shown in [3] and [7], the case of a vanishing derivative at $x = 0$ is of great interest (Child–Langmuir regime). In this case, we prove that the boundary-value problem (55)–(60), with the additional requirement that $\frac{d\varphi}{dx}(0) = 0$ and with unprescribed j, has a unique solution. We begin with the following lemma.

LEMMA 5.2. *Let $\widetilde{\varphi}$ be the solution of* (71) *with $\beta = 0$. Then $\widetilde{\varphi}$ has the following asymptotic behavior:*

$$\frac{d\widetilde{\varphi}}{dx} \sim 2\sqrt{x}, \quad \widetilde{\varphi} \sim \frac{4}{3}x^{3/2}, \qquad x \to +\infty, \tag{184}$$

$$\frac{d\widetilde{\varphi}}{dx} \sim 4/3\left(\frac{9}{4}\right)^{2/3} x^{1/3}, \quad \widetilde{\varphi} \sim \left(\frac{9}{4}\right)^{2/3} x^{4/3}, \qquad x \to 0. \tag{185}$$

Besides, $\widetilde{\varphi}$ satisfies the following estimates on $\mathbb{R}_*^+$:

$$\widetilde{\varphi}(x) > \frac{4}{3}x^{3/2}, \quad \widetilde{\varphi}(x) > \left(\frac{9}{4}\right)^{2/3} x^{4/3}. \tag{186}$$

Proof. The asymptotic behavior in the vicinity of 0 (185) is already proven in Theorem 4.1. The global estimates (186) are direct consequences of (176). Indeed, Proposition 5.1 applies for $\widetilde{\varphi}$ with $\beta_1 = 0$ and j and τ replaced by 1. Therefore, we deduce from (182) that $\widetilde{\varphi}'' > \frac{1}{\sqrt{\widetilde{\varphi}}}$. The integration of this inequality leads to the second inequality of (186). The first inequality comes from (180) applied with $\tau = j = 1$.

The only thing left to show is (184). First, we deduce from (176) that

$$\begin{aligned} 4x < \frac{d\widetilde{\varphi}}{dx}^2(x) &\le 4x + 4\sqrt{\widetilde{\varphi}(x)}g_{\widetilde{\varphi}}(x,0) + 4\sqrt{\widetilde{\varphi}(x)}\int_0^x \widetilde{n}(y)g_{\widetilde{\varphi}}(x,y)dy \\ &< 4x + 4\sqrt{\widetilde{\varphi}(x)}; \end{aligned} \tag{187}$$

therefore it is sufficient to prove that $\widetilde{\varphi} = o(x^2)$ in the neighborhood of $+\infty$. Let us introduce the function

$$f(x) = \frac{\widetilde{\varphi}(x)}{x^{5/3}}.$$

We prove now that f is decreasing in the neighborhood of $+\infty$ and consequently it is bounded. The derivative of f has the same sign as

$$A(x) = x\,\widetilde{\varphi}'(x) - \frac{5}{3}\widetilde{\varphi}(x) = x\left[\widetilde{\varphi}'(x) - \frac{3}{2}\widetilde{\varphi}(x)\right] - \frac{1}{6}\widetilde{\varphi}(x). \tag{188}$$

We deduce from (187) that $\widetilde{\varphi}'(x) \le 2\sqrt{x} + 2\widetilde{\varphi}^{1/4}(x)$, and by using the first estimate (186), we bound $A(x)$ by

$$\begin{aligned} A(x) &\le 2x\,\widetilde{\varphi}^{1/4}(x) - \frac{1}{6}\widetilde{\varphi}(x) \\ &\le \widetilde{\varphi}^{1/4}(x)\left(2x - \frac{1}{6}\widetilde{\varphi}^{3/4}(x)\right). \end{aligned}$$

Since $\widetilde{\varphi}(x) \geq \frac{4}{3}x^{3/2}$, then A is negative for large x's. This ends the proof of the lemma. □

Proof of Theorem 2.1. We set $X = \frac{1}{\tau^3 j}$. Then j is determined by the equation

$$\widetilde{\varphi}(X) = \tau^2 X^2, \tag{189}$$

where $\widetilde{\varphi}$ is given by (173). From (181) and (183), we deduce the following estimates on X:

$$X > \frac{9}{4\tau^3}, \qquad X > \frac{16}{9\tau^4}. \tag{190}$$

Therefore, we have studied the equation $B(x) := \widetilde{\varphi}(x) - \tau^2 x^2 = 0$ on the interval $(X_{\min}, +\infty)$, where

$$X_{\min} = \sup\left(\frac{9}{4\tau^3}, \frac{16}{9\tau^4}\right).$$

Thanks to the preceding lemma, it is obvious that $\lim_{x\to+\infty} B(x) = -\infty$. Besides, by inserting the inequality $\widetilde{\varphi}(x) \geq \frac{4}{3}x^{3/2}$ into the expression of B, we find

$$B\left(\frac{16}{9\tau^4}\right) \geq 0.$$

This insures the existence of a solution of $B(x) = 0$. To prove the uniqueness, it is sufficient to prove the implication

$$(x > X_{\min},\ B(x) \leq 0) \Rightarrow (B'(x) < 0). \tag{191}$$

Let us prove (191). The condition $B(x) \leq 0$ implies that $\widetilde{\varphi}(x) \leq \tau^2 x^2$. By using this inequality and (187), we obtain

$$B'(x) < 2\sqrt{1+\tau}\,\sqrt{x} - 2\tau^2 x = 2\sqrt{x}\,(\sqrt{1+\tau} - \tau^2\sqrt{x}).$$

The condition $x > \frac{16}{9\tau^4}$ leads to

$$B'(x) < 2\sqrt{x}\left(\sqrt{1+\tau} - \frac{4}{3}\right).$$

The right-hand side of this inequality is negative when $\tau \leq 7/9$, which proves (191). By using the condition $x > \frac{9}{4\tau^3}$ and proceeding in a likewise manner, we prove (191) in the case $\tau \geq 4/5$.

Finally, we prove that the current is equivalent to $\frac{9}{16}\tau$ when τ tends to zero and to $4/9$ when τ tends to infinity. Since $X(\tau) = 1/(\tau^3 j_{CL}(\tau))$ satisfies $\frac{\widetilde{\varphi}(X)}{X^2} = \tau^2$, then we deduce from the behavior of $\frac{\widetilde{\varphi}(x)}{x^2}$ that $\lim_{\tau\to 0} X(\tau) = +\infty$ and $\lim_{\tau\to\infty} X(\tau) = 0$. Thus, by using the asymptotic behavior of $\widetilde{\varphi}$ in the vicinity of zero and infinity, we obtain the result.

6. Conclusion. We derived a limit model for electron transport in a semiconductor via a Child-Langmuir asymptotics of the Boltzmann equation. The results of Theorem 2.1 seem to be true for all the values of τ but would need more technical arguments. In our analysis we neglected the doping density in the $N-$ region. One of the effects of the doping profile is the loss of the convexity of the potential which

in turn would not be necessarily increasing, and the explicit formulas for the density would not be valid. However, in the collisionless case, when the doping density is not too large, the limit potential is increasing (see [19] or [20]). This result could be extended to the case $\tau \neq \infty$. Finally, the convergence problem of the perturbation problem solutions (19)–(24) to the limit problem solution is not solved yet and its proof is in progress.

REFERENCES

[1] I. LANGMUIR AND K. T. COMPTON, *Electrical discharges in gases: Part* II, *fundamental phenomena in electrical discharges*, Rev. Modern Phys., 3 (1931), pp. 191–257.

[2] C. GREENGARD AND P. A. RAVIART, *A boundary value problem for the stationary Vlasov-Poisson equations: The plane diode*, Comm. Pure Appl. Math., 43 (1990), pp. 473–507.

[3] P. DEGOND AND P. A. RAVIART, *An asymptotic analysis of the one-dimensional Vlasov–Poisson system: the Child–Langmuir law*, Asymptotic Anal., 4 (1991), pp. 187–214.

[4] ———, *On a penalization of the Child-Langmuir emission condition for the one-dimensional Vlasov–Poisson equation*, Asymptotic Anal., 6 (1992), pp. 1–27.

[5] F. POUPAUD, *Boundary value problems for the stationary Vlasov-Maxwell systems*, Forum Mathematicum, 4 (1992), pp. 499–527.

[6] ———, *Solutions stationnaires des equations de Vlasov–Poisson*, C.R. Acad. Sci. Paris, 311 (1990), pp. 307–312.

[7] P. DEGOND, S. JAFFARD, F. POUPAUD, AND P. A. RAVIART, *The Child-Langmuir asymptotics of the Vlasov–Poisson equation for cylindrically or spherically symmetric diodes, Part* I *Statement of the problem and basic estimates, Part* II, *Analysis of the reduced problem and determination of the Child-Langmuir current*, Math. Meth. Appl. Sci., to appear.

[8] M. S. SHUR AND L. F. EASTMAN, *Ballistic transport in semiconductors at low temperatures for low-power high-speed logic*, IEEE Trans. Electron Dev., ED-26 (1979), pp. 1677–1683.

[9] ———, *Near ballistic transport in GaAs Devices at 77 K*, Solid-State Electron. 24, pp. 11–18.

[10] N. BEN ABDALLAH AND P. DEGOND, *On the Child-Langmuir law for semiconductors*, I.M.A. proceeding volume on Semiconductors, to appear.

[11] H. U. BARANGER AND J. W. WILKINS, *Ballistic electrons in a submicron structure, the distribution function and two-valley effects*, Physica B, 134 (1985), pp. 470–474.

[12] L. REGGIANI, ED., *Hot-Electron Transport in Semiconductors*, Springer-Verlag, Berlin, 1985.

[13] B. NICLOT, P. DEGOND, AND F. POUPAUD, *Deterministic particle simulations of the Boltzmann transport equation of semiconductors*, J. Comput. Phys., 78 (1988), pp. 313–349.

[14] P. DEGOND AND F. J. MUSTIELIS, *A deterministic particle method for the kinetic model of semiconductors: the homogeneous field model*, Solid-State Electron, to appear.

[15] H. U. BARANGER, *Ballistic electrons in a submicron semiconducting structure: A Boltzmann equation approach*, Ph.D. Thesis, Cornell University, Ithaca, NY, January 1986, p 40.

[16] H. U. BARANGER AND J. W. WILKINS, *Ballistic structure in the electron distribution function of small semiconducting structures: General features and specific trends*, Phys. Rev. B, 36 (1987), pp. 1487–1502.

[17] P. DEGOND AND F. GUYOT-DELAURENS, *Particle Simulation of the Semiconductor Boltzmann Equation for One-Dimensional Inhomogeneous Structures*, J. Comput. Phys., 90 (1990), pp. 65–97.

[18] F. DELAURENS AND F. J. MUSTIELIS, *A new deterministic particle method for solving kinetic transport equations: The semiconductor Boltzmann equation case*, SIAM J. Appl. Math., 52 (1992), pp. 973–988.

[19] N. BEN ABDALLAH, P. DEGOND, AND C. SCHMEISER, *On a mathematical model for hot carrier injection in semiconductors*, Math. Meth. in the Appl. Sci., 4 (1994), pp. 409–438.

[20] N. BEN ABDALLAH, *The Child-Langmuir regime for electron transport in a plasma including a background of positive ions*, Math. Models Methods Appl. Sci., 4 (1994), pp. 409–438.

[21] S. M. SZE, *Physics of Semiconductor Devices*, J. Wiley and Sons, New York, 1981.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 399–414, March 1995

GLOBAL EXISTENCE OF SOLUTIONS TO REACTION-HYPERBOLIC SYSTEMS IN ONE SPACE DIMENSION*

DANIELLE D. CARR†

Abstract. Global existence theorems to the initial value and initial-boundary value problems are proved for a general class of reaction-hyperbolic systems that arise from the transport of chemically reacting materials. The basic technique of proof is to use the geometrical properties of the equilibrium set of the nonlinear source terms to construct L^p estimates for each p and to pass to the limit to get an a priori L^∞ bound.

Key words. material transport, energy norms, chemical kinetics, hyperbolic equations, initial value problem, initial-boundary value problem

AMS subject classifications. 35A05, 35B45, 35L45, 35L60

1. Introduction.

We refer to a system of partial differential equations as reaction-hyperbolic if it has the form

$$L\vec{u}(x,t) = \vec{f}(x,t,\vec{u}),$$

where $\vec{u}(x,t)$ is an r-dimensional vector whose components represent the concentrations of the chemical species u_i, L is a smooth first-order hyperbolic operator, and the linear or nonlinear functions in $\vec{f}$ represent the chemical reactions taking place among the species u_i [9]. These systems arise naturally in the study of material transport such as the transport of intracellular materials in nerve axons [1], [3], [5]. In this paper we concentrate on a class of reaction-hyperbolic systems in one space dimension that can be cast into the canonical form

$$\begin{aligned}
(\partial_t + \lambda_1\partial_x)u_1(x,t) &= f_1(u_1,u_2),\\
&\vdots \qquad\qquad \vdots\\
(\partial_t + \lambda_i\partial_x)u_i(x,t) &= -f_{i-1}(u_{i-1},u_i) + f_i(u_i,u_{i+1}),\\
&\vdots \qquad\qquad \vdots\\
(\partial_t + \lambda_r\partial_x)u_r(x,t) &= -f_{r-1}(u_{r-1},u_r),\\
\vec{u}(x,0) &= \vec{u}_0(x),
\end{aligned} \tag{1}$$

where the initial data are nonnegative, continuously differentiable functions that vanish at infinity. Without loss of generality, we assume that all the λ_i are constant and not necessarily distinct. In fact, in many applications several of the λ_i are zero [2].

The problem is said to have a global solution if for any $T \in \mathbb{R}^+$ a bounded solution exists for all time $t \in [0,T]$. Since we allow the source terms to be nonlinear, global existence to system (1) is not obvious. Consider the case where $\vec{u} = (u_1(x,t), u_2(x,t))$ for $x \in \mathbb{R}$. For fixed $\lambda_1, \lambda_2 \in \mathbb{R}$, system (1) reduces to

$$\begin{aligned}
(\partial_t + \lambda_1\partial_x)u_1(x,t) &= f(\vec{u}),\\
(\partial_t + \lambda_2\partial_x)u_2(x,t) &= -f(\vec{u}),\\
\vec{u}(x,0) &= \vec{u}_0(x),
\end{aligned} \tag{2}$$

* Received by the editors April 23, 1993; accepted for publication October 25, 1993. This work was supported in part by the National Science Foundation Minority Graduate Fellowship award and the National Science Foundation Postdoctoral Fellowship award.

† Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York, 10012 (dcarr@cims.nyu.edu).

which describes the chemical interactions and the active transport of the species u_i at the rate λ_i along the real line. Note that for $f(\vec{u}) \equiv u_1^2$ and positive initial data, the system

$$\begin{aligned}(\partial_t + \lambda_1\partial_x)u_1(x,t) &= u_1^2,\\ (\partial_t + \lambda_2\partial_x)u_2(x,t) &= -u_1^2,\\ \vec{u}(x,0) &= \vec{u}_0(x),\end{aligned}$$

has no global solutions, since the first characteristic equation $\dot{u}_1 = u_1^2$ has no global solution for positive initial values.

We establish conditions on the source terms so that system (2) and ultimately system (1) are globally solvable. We do this by constructing an interesting family of energy norms which depends in an explicit way on the geometry of the equilibrium set of the source terms. These norms allow us to derive L^p estimates for the solution for each p, and taking the limit we obtain an L^∞ estimate. Global existence follows easily.

In §2, we exploit the fundamental properties of chemical kinetics in order to prove that a local solution exists to system (1) and that if the local solution starts positive for all $x \in \mathbb{R}$ then it stays positive for all local time. If the local solution can be majorized by a constant C that depends only on the initial data and on a fixed time step, then the solution exists globally. This is achieved by repeated successive approximations along the time axis, using the fact that the solution remains bounded by C. In §3, we establish conditions on the source terms for the 2×2 system in order to construct an a priori bound for the solution to the initial value problem. This a priori estimate depends not only on the initial data, but also on the geometry of the equilibrium set of the nonlinear source terms f_i. A special case of the 2×2 system extends Illner's results [6]. In §4, we generalize our results to certain classes of $r \times r$ systems, and finally in §5 we indicate how these results are extended to the initial-boundary value problem.

There are lots of important equations without the form (1), for which one can prove global existence using other methods. Some of these techniques are clearly illustrated in the proof of global existence to the fast axonal transport system [5].

2. Local properties. Chemical reactions can be classified on a kinetic basis by reaction order [7]. For $p > 0$, pth-order reactions are those whose rate is proportional to the product of the concentrations of p reactants. Let $\vec{u} = (u_1(x,t), \ldots, u_r(x,t))$ denote the concentrations of all the chemical species u_i used and produced in a chemical system at position x and at time t. The reaction

$$u_1 + u_2 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} u_3$$

is second order since the formation of u_3 depends on the concentrations of two reactants, u_1 and u_2. The resulting rate equations for the concentrations of u_1, u_2, and u_3 are

$$\begin{aligned}\partial_t u_1(x,t) &= -k_1u_1u_2 + k_2u_3,\\ \partial_t u_2(x,t) &= -k_1u_1u_2 + k_2u_3,\\ \partial_t u_3(x,t) &= k_1u_1u_2 - k_2u_3.\end{aligned}$$

Note that $\partial_t(u_1 + u_2 + 2u_3) = 0$, which is just the conservation of mass.

In general, for a closed system the equations that describe the rates of change of the chemical concentrations u_i with respect to time are

$$\begin{aligned} \partial_t u_i(x,t) &= f_i(\vec{u}), \\ \vec{u}(x,0) &= \vec{u}_0(x), \end{aligned} \tag{3}$$

where $i = 1, \ldots, r$ and the linear or nonlinear functions f_i are rate functions which represent the chemical reactions involving the chemical species u_i. Note that if p is the maximum order of the chemical reactions, then each f_i is a kth-order polynomial of $\vec{u}$ for $0 < k \leq p$. In addition, since the loss of a certain species is proportional to its present amount, u_i is a factor in all terms with negative coefficients in the polynomial f_i. Since the chemical system is closed, mass is conserved and the net reaction rate is zero. Consequently, there exists a linear combination of the f_i involving positive constants such that the sum is zero. Finally, if all the concentrations of the chemical species are zero, then the values of the rate functions f_i are zero. These fundamental properties motivate the following structural hypotheses on the source terms f_i.

DEFINITION 2.1. *$\vec{f}$ is a* reaction function *if and only if $\vec{f}$ satisfies the following three properties:*

(i) *For $p \geq 1$, the components of $\vec{f}$ have the following form:*

$$f_i(x,t,\vec{u}) = \sum_{|\alpha^i| \leq p} c_{\alpha^i}(x,t,\vec{u})\vec{u}^{\alpha^i} - u_i \sum_{|\beta^i| \leq (p-1)} d_{\beta^i}(x,t,\vec{u})\vec{u}^{\beta^i}, \tag{4}$$

where for each $i = 1, \ldots, r$ and $j = 1, \ldots, r$, α_j^i and β_j^i are nonnegative constants; the coefficients $c_{\alpha_j^i}(x,t,\vec{u})$ and $d_{\beta_j^i}(x,t,\vec{u})$ are smooth, nonnegative, uniformly bounded functions in x and t.

(ii) *There exist positive constants b_i such that $\sum_{i=1}^r b_i f_i = 0$.*

(iii) $\vec{f}(x,t,\vec{0}) = \vec{0}$.

Let $C_\infty(\mathbb{R})$ denote the set of all real-valued, continuous functions vanishing at infinity on $\mathbb{R}$. $C_\infty([0,\delta) \times \mathbb{R})$ is the set of all real-valued functions w such that for all $\epsilon > 0$ there is a closed interval $I \subset \mathbb{R}$ with $\mid w(x,t) \mid < \epsilon$ for all $t \in [0,\delta)$ and all $x \notin I$. If we write C_∞^1, we mean the set of those functions, which are continuously differentiable in addition to having the above properties.

PROPOSITION 2.2 (local existence and positivity). *Consider the following system:*

$$\begin{aligned} (\partial_t + \lambda_i \partial_x) u_i(x,t) &= f_i(x,t,\vec{u}), \\ \vec{u}(x,0) &= \vec{u}_0(x), \end{aligned} \tag{5}$$

where $\vec{f} \equiv (f_1, \ldots, f_r)$ is a reaction function, and the initial data u_{0i} are uniformly bounded, nonnegative, $C_\infty^1(\mathbb{R})$ with uniformly bounded derivatives. There is a $\delta > 0$ so that system (5) has a unique nonnegative solution in C_∞^1 for $t \in [0,\delta)$.

Proof. The proof is a straightforward generalization of Cabannes' result, using the method of successive approximations [4]. For details, see [5]. □

3. 2 × 2 systems. Even though we know the basic structure of the reaction function $\vec{f}$, global existence to system (1) is still not obvious since we have not addressed its highly nonlinear nature. In order to figure out the right additional hypotheses for the source terms, consider the following simple chemical equation:

$$\mu u_1 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \eta u_2, \tag{6}$$

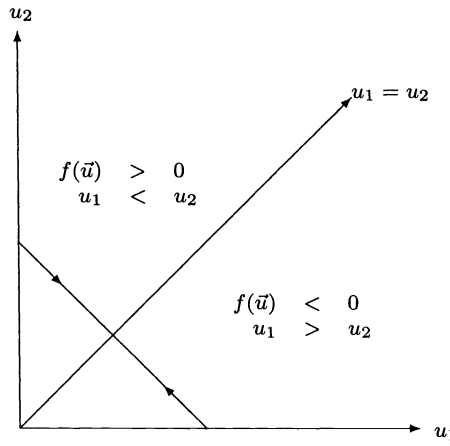


FIG. 1. *Stable equilibria for* $f(\vec{u}) = -k_1 u_1^\eta + k_1 u_2^\eta$.

which leads to the following differential equations for the spatially homogeneous case:

$$\begin{aligned} \partial_t u_1(x,t) &= -k_1 u_1^\mu + k_2 u_2^\eta, \\ \partial_t u_2(x,t) &= k_1 u_1^\mu - k_2 u_2^\eta, \end{aligned} \tag{7}$$

where $k_1, k_2, \eta, \mu > 0$. For $\eta = \mu$ and $k_1 = k_2$, (7) reduces to

$$\begin{aligned} \partial_t u_1(x,t) &= -k_1 u_1^\eta + k_1 u_2^\eta, \\ \partial_t u_2(x,t) &= k_1 u_1^\eta - k_1 u_2^\eta. \end{aligned} \tag{8}$$

Let $f(u_1, u_2) \equiv -k_1 u_1^\eta + k_1 u_2^\eta$ and let $\vec{f}(\vec{u}) \equiv (f(\vec{u}), -f(\vec{u}))$. Note that the chemical system (6) is at steady state when the components of the source term are zero and that this occurs only when the concentrations of the reactants are equal. Define the zero curve for f to be $z(\tau) \equiv \tau$. If the concentration of u_1 is initially less than $z(u_2)$, then f is strictly positive. In this case, the concentration of u_1 increases while the concentration of u_2 decreases until the two are equal. Similarly, if the concentration of u_1 is initially greater than $z(u_2)$, then f is strictly negative, and the chemistry adjusts itself until the two concentrations are equal (see Fig. 1). Thus, the zero line $z(\cdot)$ consists of stable equilibria for system (8). One can easily see that $\max\{\| u_1(x,t) \|_\infty, \| u_2(x,t) \|_\infty\}$ is nonincreasing in time and that the maximum of the initial data is the a priori bound for the solution [6].

For the case when the zero curve no longer coincides with the line $u_1 = u_2$, this bound does not hold. Without loss of generality, let $\frac{k_2}{k_1} = 2$. The zero line $z(u_2) \equiv 2u_2$ still consists of stable equilibria for (7); however, the $\max\{\| u_1(x,t) \|_\infty, \| u_2(x,t) \|_\infty\}$ increases in time in the wedge $u_2 < u_1$ and $u_2 > \frac{1}{2}u_1$ (see Fig. 2). New estimates must be constructed which utilize the highly stable structure of the spatially homogeneous case.

DEFINITION 3.1. *We define a class of source terms* $\vec{f}$ *such that system* (2) *has a curve of stable equilibria for the spatially homogeneous case. We say that* $\vec{f}$ *satisfies the* stability criteria *for the* 2×2 *system if and only if*

(i) $\vec{f}(\vec{u}) \equiv (f(\vec{u}), -f(\vec{u}))$ *is a* C^2 *reaction function of* $\vec{u} \equiv (u_1, u_2)$ *only.*

(ii) *The set of zeros of* f *in the closed first quadrant is given by a continuous, one-to-one function,* $z(\cdot)$, *where* $z(0) \equiv 0$.

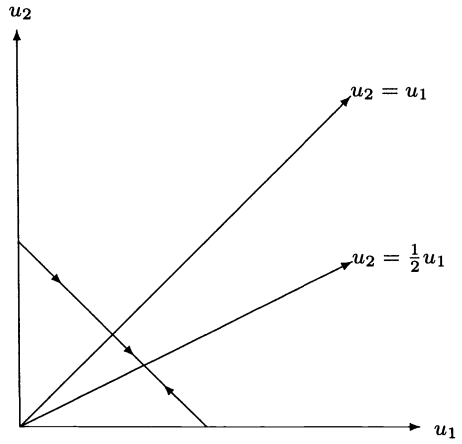


FIG. 2. *Stable equilibria for* $f(\vec{u}) = -k_1 u_1^\eta + k_2 u_2^\eta$, *where* $\frac{k_2}{k_1} = 2$.

(iii) $f(u_1, u_2)(u_1 - z(u_2)) \leq 0$ *for all* $x \in \mathbb{R}$ *and* $t \in [0, \delta)$.

We now prove that if the source term $\vec{f}$ in (2) satisfies the stability criteria, then a global solution exists to (2).

THEOREM 3.2. *Consider the* 2×2 *system:*

$$
\begin{aligned}
(\partial_t + \lambda_1 \partial_x) u_1(x,t) &= f(\vec{u}), \\
(\partial_t + \lambda_2 \partial_x) u_2(x,t) &= -f(\vec{u}), \\
\vec{u}(x,0) &= \vec{u}_0(x),
\end{aligned} \tag{9}
$$

where λ_1 *and* λ_2 *are constants and the initial data* u_{0i} *are uniformly bounded, nonnegative,* $C^1(\mathbb{R}) \cap L^1(\mathbb{R})$ *functions with uniformly bounded derivatives. If the source term* $\vec{f}(\vec{u}) \equiv (f(\vec{u}), -f(\vec{u}))$ *satisfies the stability criteria, then a solution to* (9) *exists for all* $x \in \mathbb{R}$ *and for all* $t \geq 0$.

Proof. To prove Theorem 3.2, we derive an L^∞ a priori estimate for the solution for all time.

By Proposition 2.2, a nonnegative C^1_∞ local solution exists to (9). Pick $T > 0$ and assume that a nonnegative solution exists for all $x \in \mathbb{R}$ and $t \in [0, T]$. For $n > 1$, define the energy norm E_{u_1} as

$$
E_{u_1}(t) \equiv \int_{-\infty}^{\infty} \left[u_1^n(x,t) + \int_0^{u_2(x,t)} n z^{n-1}(\tau) d\tau \right] dx. \tag{10}
$$

All terms in (10) are nonnegative by Proposition 2.2 and the stability criteria. Thus, for all $t \in [0, T]$ we know that

$$
E_{u_1}(t) \geq \int_{-\infty}^{\infty} u_1^n(x,t) dx.
$$

Differentiating E_{u_1} with respect to time and using the differential equations yields the following equalities:

$$
\frac{d}{dt} E_{u_1}(t) = \int_{-\infty}^{\infty} \left[n u_1^{n-1}(x,t) \partial_t u_1(x,t) + n z^{n-1}(u_2) \partial_t u_2(x,t) \right] dx
$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \left[nf(\vec{u})(u_1^{n-1}(x,t) - z^{n-1}(u_2)) \right. \\
&\qquad \left. -\lambda_1 n u_1^{n-1} \partial_x u_1(x,t) - \lambda_2 n z^{n-1}(u_2) \partial_x u_2(x,t) \right] dx \\
&= \int_{-\infty}^{\infty} nf(\vec{u})(u_1^{n-1}(x,t) - z^{n-1}(u_2))dx \\
&\quad - \int_{-\infty}^{\infty} \lambda_1 \partial_x [u_1^n(x,t)]dx - \int_{-\infty}^{\infty} \lambda_2 \partial_x \int_0^{u_2(x,t)} n z^{n-1}(\tau) d\tau dx \\
&= \int_{-\infty}^{\infty} nf(\vec{u})(u_1^{n-1}(x,t) - z^{n-1}(u_2))dx \\
&\quad -\lambda_1 \lim_{x\to\infty} u_1^n(x,t) + \lambda_1 \lim_{x\to-\infty} u_1^n(x,t) \qquad (11) \\
&\quad -\lambda_2 \lim_{x\to\infty} \int_0^{u_2(x,t)} n z^{n-1}(\tau) d\tau + \lambda_2 \lim_{x\to-\infty} \int_0^{u_2(x,t)} n z^{n-1}(\tau) d\tau.
\end{aligned}
$$

All the indicated limits in (11) are zero since the initial data are in $C^1_\infty(\mathbb{R})$. By the stability criteria we know that for all nonnegative u_i, $f(u_1, u_2)(u_1^{n-1} - z^{n-1}(u_2)) \leq 0$. We thus have

$$
\frac{d}{dt} E_{u_1}(t) = \int_{-\infty}^{\infty} nf(\vec{u})(u_1^{n-1}(x,t) - z^{n-1}(u_2))dx \quad \leq \quad 0.
$$

The energy norm E_{u_1} was carefully constructed to be a nonincreasing function of time. Thus, for $0 \leq t \leq T$

$$
E_{u_1}(0) \geq E_{u_1}(t) \geq \int_{-\infty}^{\infty} u_1^n(x,t)dx. \qquad (12)
$$

Estimating $E_{u_1}(0)$ yields the following:

$$
\begin{aligned}
E_{u_1}(0) &= \int_{-\infty}^{\infty} \left[u_{01}^n(x) + \int_0^{u_{02}(x)} n z^{n-1}(\tau) d\tau \right] dx \\
&\leq \int_{-\infty}^{\infty} u_{01}^n(x)dx + n \left(\max_{0\leq s\leq \| u_{02} \|_\infty} z(s) \right)^{n-1} \int_{-\infty}^{\infty} u_{02}(x)dx \\
&\leq \| u_{01} \|_\infty^{n-1} \| u_{01} \|_1 + nz(\| u_{02} \|_\infty)^{n-1} \| u_{02}(x) \|_1 . \qquad (13)
\end{aligned}
$$

Combining (12) and (13) yields the following estimate:

$$
\begin{aligned}
\left\{ \int_{-\infty}^{\infty} u_1^n(x,t)dx \right\}^{\frac{1}{n}} &\leq \{ \| u_{01} \|_\infty^{n-1} \| u_{01} \|_1 \}^{\frac{1}{n}} \qquad (14) \\
&\quad + \{ nz(\| u_{02} \|_\infty)^{n-1} \| u_{02}(x) \|_1 \}^{\frac{1}{n}}.
\end{aligned}
$$

We know that for a given function w if $\overline{\lim}_{p\to\infty} \| w \|_p = s < \infty$, then $\| w \|_\infty \leq s$. Taking the lim sup of both sides of (14), we get the desired L^∞ estimate

$$
\| u_1(x,t) \|_\infty \leq \| u_{01}(x) \|_\infty + z(\| u_{02} \|_\infty) \qquad (15)
$$

for all $t \in [0, T]$.

To find an a priori bound for u_2, we apply a similar argument. Let $m > 1$ and denote the inverse function of $z(\cdot)$ as $\gamma(\cdot)$. Define the energy norm E_{u_2} as

$$E_{u_2}(t) \equiv \int_{-\infty}^{\infty} \left[u_2^m(x,t) + \int_0^{u_1(x,t)} m\gamma^{m-1}(\tau)d\tau \right] dx. \tag{16}$$

Differentiating E_{u_2} with respect to time yields

$$\begin{aligned}
\frac{d}{dt}E_{u_2}(t) &= \int_{-\infty}^{\infty} \left[mu_2^{m-1}(x,t)\partial_t u_2(x,t) + m\gamma^{m-1}(u_1)\partial_t u_1(x,t) \right] dx \\
&= \int_{-\infty}^{\infty} \left[mf(\vec{u})(\gamma^{m-1}(u_1) - u_2^{m-1}(x,t)) \right. \\
&\qquad \left. -\lambda_2 m u_2^{m-1}\partial_x u_2(x,t) - \lambda_1 m\gamma^{m-1}(u_1)\partial_x u_1(x,t) \right] dx \\
&= \int_{-\infty}^{\infty} mf(\vec{u})(\gamma^{m-1}(u_1) - u_2^{m-1}(x,t))dx \\
&\quad -\lambda_2 \lim_{x\to\infty} u_2^m(x,t) + \lambda_2 \lim_{x\to-\infty} u_2^m(x,t) \\
&\quad -\lambda_1 \lim_{x\to\infty} \int_0^{u_1(x,t)} m\gamma^{m-1}(\tau)d\tau \\
&\quad + \lambda_1 \lim_{x\to-\infty} \int_0^{u_1(x,t)} m\gamma^{m-1}(\tau)d\tau \\
&= \int_{-\infty}^{\infty} mf(\vec{u})(\gamma^{m-1}(u_1) - u_2^{m-1}(x,t))dx \quad \leq 0
\end{aligned} \tag{17}$$

by Proposition 2.2 and the stability criteria. Since E_{u_2} is nonincreasing in time, we have by positivity that

$$E_{u_2}(0) \geq E_{u_2}(t) \geq \int_{-\infty}^{\infty} u_2^m(x,t)dx \tag{18}$$

for $0 \leq t \leq T$. Applying the previous argument to estimate $E_{u_2}(0)$, we get our final result for all $t \in [0,T]$:

$$\| u_2(x,t) \|_\infty \leq \| u_{02}(x) \|_\infty + \gamma(\| u_{01} \|_\infty). \tag{19}$$

We have shown that $\max\{\| u_1(x,t) \|_\infty, \| u_2(x,t) \|_\infty\} \leq \mathcal{A}$ for all $t \in [0,T]$, where the a priori bound is defined as

$$\mathcal{A} \equiv \max\{\| u_{01} \|_\infty + z(\| u_{02} \|_\infty), \| u_{02} \|_\infty + \gamma(\| u_{01} \|_\infty)\}. \tag{20}$$

This a priori bound ensures that a solution to (9) exists for all time and space. □

Note that the a priori estimate (20) depends not only on the initial data, but also on the geometry of the zero set of the source terms. As an example, consider the system

$$\begin{aligned}
(\partial_t + \lambda_1\partial_x)u_1(x,t) &= -k_1 u_1^\mu + k_2 u_2^\eta, \\
(\partial_t + \lambda_2\partial_x)u_2(x,t) &= k_1 u_1^\mu - k_2 u_2^\eta, \\
\vec{u}(x,0) &= \vec{u}_0(x),
\end{aligned} \tag{21}$$

where η, μ, k_1, and k_2 are positive constants, λ_1 and λ_2 are arbitrary constants, and for $i = 1,2$ the initial data u_{0i} are nonnegative, uniformly bounded, $C^1(\mathbb{R}) \cap L^1(\mathbb{R})$

functions with uniformly bounded derivatives. The curve of zeros is simply $z(u_2) = (\frac{k_2}{k_1})^{1/\mu} u_2^{\eta/\mu}$, and it is clear that $\vec{f}(\vec{u}) \equiv (-k_1 u_1^\mu + k_2 u_2^\eta, k_1 u_1^\mu - k_2 u_2^\eta)$ satisfies the stability criteria. Thus, by Theorem 3.2 a global solution exists to (21) for any $\eta > 0$ and $\mu > 0$.

Let $\mathcal{K}_1 = (\frac{k_2}{k_1})^{1/\mu}$, $\mathcal{K}_2 = (\frac{k_1}{k_2})^{1/\eta}$, and $\rho = \frac{\eta}{\mu}$. By (20), the a priori bound for the solution to system (21) is

$$\mathcal{A} \equiv \max\{\| u_{01}(x) \|_\infty + \mathcal{K}_1 \| u_{02}(x) \|_\infty^\rho, \mathcal{K}_2 \| u_{01}(x) \|_\infty^{\frac{1}{\rho}} + \| u_{02}(x) \|_\infty\}.$$

Now consider the case when $\mu = \eta$. For $\mathcal{K} = (\frac{k_2}{k_1})^{1/\eta}$, the a priori bound reduces to

$$\mathcal{A} \equiv \max\{\| u_{01}(x) \|_\infty + \mathcal{K} \| u_{02}(x) \|_\infty, \mathcal{K}^{-1} \| u_{01}(x) \|_\infty + \| u_{02}(x) \|_\infty\}.$$

In addition if we set $k_1 = k_2$, the energy norms are simply

$$E_{u_1} = \int_{-\infty}^{\infty} [u_1^n(x,t) + u_2^n(x,t)]\, dx,$$

$$E_{u_2} = \int_{-\infty}^{\infty} [u_2^m(x,t) + u_1^m(x,t)]\, dx,$$

and the a priori bound $\mathcal{A}$ further reduces to

$$\mathcal{A} \equiv \| u_{01}(x) \|_\infty + \| u_{02}(x) \|_\infty .$$

We remark that Illner showed global existence in this case for $\eta = 2$ in [6].

There are various degenerate systems that can be handled by different yet very simple methods. For $k_1, \mu > 0$ the source term in the system

$$\begin{aligned} (\partial_t + \lambda_1 \partial_x) u_1(x,t) &= -k_1 u_1^\mu, \\ (\partial_t + \lambda_2 \partial_x) u_2(x,t) &= k_1 u_1^\mu, \\ u_i(x,0) &= u_{0i}(x), \end{aligned} \tag{22}$$

does not satisfy the stability criteria since the zero curve is not invertible. System (22) can be written as

$$\begin{aligned} (\partial_t + \lambda_1 \partial_x) u_1(x,t) &= f(u_1), \\ (\partial_t + \lambda_2 \partial_x) u_2(x,t) &= -f(u_1), \\ u_i(x,0) &= u_{0i}(x), \end{aligned} \tag{23}$$

where f has the following form:

$$f(u_1) \equiv -u_1 \sum_{|\beta_1^i| \le \mu - 1} d_{\beta_1^i}(x,t,u_1) u_1^{\beta_1^i}$$

(see formula (4)). Since $\vec{f}(u_1) \equiv (f(u_1), -f(u_1))$, we know by Proposition 2.2 that a nonnegative, C_∞^1 local solution exists to (23). For $T > 0$, assume that a nonnegative, C_∞^1 solution exists for $t \in [0,T]$. By positivity, we have that

$$\begin{aligned} u_1(x,t) &= u_{01}(x) + \int_0^t f(u_1(x - \lambda_1(t-s), s))ds \\ &\le \| u_{01}(x) \|_\infty . \end{aligned}$$

Since $f(u_1)$ is a smooth function on a compact set, it has a maximum in terms of the initial datum $u_{01}(x)$. Using positivity, we analyze the integral equation for u_2:

$$\begin{aligned} u_2(x,t) &= u_{02}(x) - \int_0^t f(u_1(x-\lambda_2(t-s),s))ds \\ &\leq \| u_{02}(x) \|_\infty + T \max(-f(\| u_{01}(x) \|_\infty)). \end{aligned}$$

Thus, the a priori bound for the solution to (22) for all $x \in \mathbb{R}$ and $t \in [0,T]$ is

$$\mathcal{A} \equiv \max\{\| u_{01}(x) \|_\infty, \ \| u_{02}(x) \|_\infty + Tk_1 \| u_{01}(x) \|_\infty^\mu\}.$$

4. r × r systems. We now extend the results for the 2×2 system to a certain class of $r \times r$ reaction-hyperbolic systems with the following form:

$$\begin{aligned} (\partial_t + \lambda_1 \partial_x) u_1(x,t) &= f_1(u_1,u_2), \\ &\vdots \qquad\qquad \vdots \\ (\partial_t + \lambda_i \partial_x) u_i(x,t) &= -f_{i-1}(u_{i-1},u_i) + f_i(u_i,u_{i+1}), \\ &\vdots \qquad\qquad \vdots \\ (\partial_t + \lambda_r \partial_x) u_r(x,t) &= -f_{r-1}(u_{r-1},u_r), \\ \vec{u}(x,0) &= \vec{u}_0(x), \end{aligned} \tag{24}$$

where $r > 2$. For $i = 1, \ldots, r$, let F_i be the right-hand side of the ith equation. We extend the definition of stability criteria to $r \times r$ systems.

DEFINITION 4.1. *For the $r \times r$ case, we say that $\vec{F} \equiv (F_1, \ldots, F_r)$ satisfies the stability criteria if and only if*

(i) *$\vec{F}(\vec{u})$ is a C^2 reaction function of $\vec{u}$ only.*

(ii) *For $k = 1, \ldots, r-1$, the set of zeros of f_k in the closed first quadrant is given by a continuous, one-to-one function, $z_k(\cdot)$, where $z_k(0) \equiv 0$.*

(iii) *For $k = 1, \ldots, r-1$, $f_k(u_k, u_{k+1})(u_k - z(u_{k+1})) \leq 0$ for all $x \in \mathbb{R}$ and $t \in [0, \delta)$.*

THEOREM 4.2. *Consider the $r \times r$ system (24), where $r > 2$. If $\vec{F} \equiv (F_1, \ldots, F_r)$ satisfies the stability criteria and the components of the initial data $\vec{u}_0(x)$ are nonnegative, uniformly bounded, $C^1(\mathbb{R}) \cap L^1(\mathbb{R})$ functions, then a global solution exists to system (24).*

Proof. Let $z_k(\cdot)$ denote the zero curve for f_k, and let $\gamma_k(\cdot)$ be its inverse. For $n > 1$, define the following functions:

$$\begin{aligned} \mathcal{Z}_{(\mathrm{m,j})}(\tau) &\equiv z_m \circ z_{m+1} \circ \cdots \circ z_j(\tau), \qquad m \leq j, \\ \Gamma_{(\mathrm{m,j})}(\tau) &\equiv \gamma_m \circ \gamma_{m-1} \circ \cdots \circ \gamma_j(\tau), \qquad m \geq j. \end{aligned} \tag{25}$$

Note that $\mathcal{Z}_{(\mathrm{m,j})}$ and $\Gamma_{(\mathrm{m,j})}$ are compositions of nonnegative, monotonic functions, and are therefore nonnegative and monotonic. Define the following energy norms:

$$E_{u_1}(t) \equiv \int_{-\infty}^{\infty} \left(u_1^n(x,t) + \sum_{k=2}^{r} \int_0^{u_k(x,t)} n \mathcal{Z}_{(1,k-1)}^{n-1}(\tau) d\tau \right) dx,$$

$$\begin{aligned} E_{u_i}(t) \equiv \int_{-\infty}^{\infty} \Bigg(u_i^n(x,t) &+ \sum_{k=1}^{i-1} \int_0^{u_k(x,t)} n \Gamma_{(\mathrm{i-1,k})}^{n-1}(\tau) d\tau \\ &+ \sum_{k=i+1}^{r} \int_0^{u_k(x,t)} n \mathcal{Z}_{(\mathrm{i,k-1})}^{n-1}(\tau) d\tau \Bigg) dx, \end{aligned}$$

$$E_{u_r}(t) \equiv \int_{-\infty}^{\infty} \left(u_r^n(x,t) + \sum_{k=1}^{r-1} \int_0^{u_k(x,t)} n\Gamma^{n-1}_{(\mathrm{r-1,k})}(\tau)d\tau \right) dx.$$

All the terms in the energy norms are nonnegative by Proposition 2.2 and (25). Differentiating $E_{u_i}(t)$ with respect to time yields the following equalities:

$$\begin{aligned}
\frac{d}{dt}E_{u_i}(t) &= \int_{-\infty}^{\infty} \left(nu_i^{n-1}(-f_{i-1} + f_i - \lambda_i \partial_x u_i) + n\Gamma^{n-1}_{(\mathrm{i-1,1})}(u_1)(f_1 - \lambda_1 \partial_x u_1) \right) dx \\
&\quad + \int_{-\infty}^{\infty} \left(\sum_{k=2}^{i-1} n\Gamma^{n-1}_{(\mathrm{i-1,k})}(u_k)(-f_{k-1} + f_k - \lambda_k \partial_x u_k) \right. \\
&\qquad \left. + \sum_{k=i+1}^{r-1} n\mathcal{Z}^{n-1}_{(\mathrm{i,k-1})}(u_k)(-f_{k-1} + f_k - \lambda_k \partial_x u_k) \right) dx \\
&\quad + \int_{-\infty}^{\infty} n\mathcal{Z}^{n-1}_{(\mathrm{i,r-1})}(u_r)(-f_{r-1} - \lambda_r \partial_x u_r) dx \\
&= \int_{-\infty}^{\infty} \left(nu_i^{n-1}(-f_{i-1} + f_i) + n\Gamma^{n-1}_{(\mathrm{i-1,1})}(u_1) f_1 \right) dx \\
&\quad + \int_{-\infty}^{\infty} \left(\sum_{k=2}^{i-1} n\Gamma^{n-1}_{(\mathrm{i-1,k})}(u_k)(-f_{k-1} - f_k) \right. \\
&\qquad \left. + \sum_{k=i+1}^{r-1} n\mathcal{Z}^{n-1}_{(\mathrm{i,k-1})}(u_k)(-f_{k-1} + f_k) \right) dx \\
&\quad + \int_{-\infty}^{\infty} n\mathcal{Z}^{n-1}_{(\mathrm{i,r-1})}(u_r)(-f_{r-1}) dx \\
&\quad + \int_{-\infty}^{\infty} \left(-\lambda_i n u_i^{n-1} \partial_x u_i - \sum_{k=1}^{i-1} n\Gamma^{n-1}_{(\mathrm{i-1,k})}(u_k)\lambda_k \partial_x u_k \right. \\
&\qquad \left. - \sum_{k=i+1}^{r} n\mathcal{Z}^{n-1}_{(\mathrm{i,k-1})}(u_k)\lambda_k \partial_x u_k \right) dx \\
&= \int_{-\infty}^{\infty} \left(nu_i^{n-1}(-f_{i-1} + f_i) + n\Gamma^{n-1}_{(\mathrm{i-1,1})}(u_1) f_1 \right) dx \\
&\quad + \int_{-\infty}^{\infty} \left(\sum_{k=2}^{i-1} n\Gamma^{n-1}_{(\mathrm{i-1,k})}(u_k)(-f_{k-1} - f_k) \right. \\
&\qquad \left. + \sum_{k=i+1}^{r-1} n\mathcal{Z}^{n-1}_{(\mathrm{i,k-1})}(u_k)(-f_{k-1} + f_k) \right) dx \qquad (26) \\
&\quad + \int_{-\infty}^{\infty} n\mathcal{Z}^{n-1}_{(\mathrm{i,r-1})}(u_r)(-f_{r-1}) dx \\
&\quad + \int_{-\infty}^{\infty} \left(-\lambda_i \partial_x [u_i^n] - \sum_{k=1}^{i-1} \partial_x \left[\int_0^{u_k(x,t)} n\lambda_k \Gamma^{n-1}_{(\mathrm{i-1,k})}(\tau)d\tau \right] \right. \qquad (27) \\
&\qquad \left. - \sum_{k=i+1}^{r} \partial_x \left[\int_0^{u_k(x,t)} n\lambda_k \mathcal{Z}^{n-1}_{(\mathrm{i,k-1})}(\tau)d\tau \right] \right) dx.
\end{aligned}$$

The boundary terms in (27) vanish since the solution is in $C^1_\infty(\mathbb{R}\times[0,T])$. We now expand the sums in (26) to get the following:

$$\begin{aligned}
\frac{d}{dt}E_{u_i} = \int_{-\infty}^{\infty} \Big(& nf_1(u_1,u_2)\left[\Gamma^{n-1}_{(\mathrm{i}-1,1)}(u_1) - \Gamma^{n-1}_{(\mathrm{i}-1,2)}(u_2)\right] \\
& + nf_2(u_2,u_3)\left[\Gamma^{n-1}_{(\mathrm{i}-1,2)}(u_2) - \Gamma^{n-1}_{(\mathrm{i}-1,3)}(u_3)\right] \\
& + \cdots + nf_{i-1}(u_{i-1},u_i)\left[\Gamma^{n-1}_{(\mathrm{i}-1,\mathrm{i}-1)}(u_{i-1}) - u_i^{n-1}\right] \\
& + nf_i(u_i,u_{i+1})\left[u_i^{n-1} - \mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{i})}(u_{i+1})\right] \\
& + nf_{i+1}(u_{i+1},u_{i+2})\left[\mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{i})}(u_{i+1}) - \mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{i}+1)}(u_{i+2})\right] \\
& + \cdots + nf_{r-1}(u_{r-1},u_r)\left[\mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{r}-2)}(u_{r-1}) - \mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{r}-1)}(u_r)\right]\Big)\,dx \\
= \int_{-\infty}^{\infty} \Big(& nf_1(u_1,u_2)\left[\Gamma^{n-1}_{(\mathrm{i}-1,2)}(\gamma_1(u_1)) - \Gamma^{n-1}_{(\mathrm{i}-1,2)}(u_2)\right] \\
& + nf_2(u_2,u_3)\left[\Gamma^{n-1}_{(\mathrm{i}-1,3)}(\gamma_2(u_2)) - \Gamma^{n-1}_{(\mathrm{i}-1,3)}(u_3)\right] \\
& + \cdots + nf_{i-1}(u_{i-1},u_i)\left[\gamma_{i-1}^{n-1}(u_{i-1}) - u_i^{n-1}\right] \\
& + nf_i(u_i,u_{i+1})\left[u_i^{n-1} - z_i^{n-1}(u_{i+1})\right] \\
& + nf_{i+1}(u_{i+1},u_{i+2})\left[\mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{i})}(u_{i+1}) - \mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{i})}(z_{i+1}(u_{i+2}))\right] \\
& + \cdots + nf_{r-1}(u_{r-1},u_r)\left[\mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{r}-2)}(u_{r-1}) - \mathcal{Z}^{n-1}_{(\mathrm{i},\mathrm{r}-2)}(z_{r-1}(u_r))\right]\Big)\,dx. \qquad (28)
\end{aligned}$$

By the stability criteria, $f_i(u_i,u_{i+1}) \geq 0$ for $\gamma_i(u_i) \leq u_{i+1}$ and for $z_i(u_{i+1}) \geq u_i$; we also have that and $f_i(u_i,u_{i+1}) \leq 0$ for $\gamma_i(u_i) \geq u_{i+1}$ and for $z_i(u_{i+1}) \leq u_i$. Since $\Gamma_{(\mathrm{m},\mathrm{j})}$ and $\mathcal{Z}_{(\mathrm{m},\mathrm{j})}$ are nonnegative monotonic functions, we have that each term in the integrand (28) is nonpositive and that

$$\frac{d}{dt}E_{u_i}(t) \leq 0.$$

E_{u_1} and E_{u_r} were carefully constructed to be nonincreasing functions of time as well. Applying the argument in Theorem 3.2 to the energy norms yields the following a priori bounds for the solution:

$$\begin{aligned}
\| u_1 \|_\infty &\leq \| u_{01} \|_\infty + \sum_{k=2}^{r} \mathcal{Z}_{(1,\mathrm{k}-1)}(\| u_{0k} \|_\infty) \\
\| u_i \|_\infty &\leq \| u_{0i} \|_\infty + \sum_{k=1}^{i-1} \Gamma_{(\mathrm{i}-1,\mathrm{k})}(\| u_{0k} \|_\infty) + \sum_{k=i+1}^{r} \mathcal{Z}_{(\mathrm{i},\mathrm{k}-1)}(\| u_{0k} \|_\infty) \qquad (29) \\
\| u_r \|_\infty &\leq \| u_{0r} \|_\infty + \sum_{k=1}^{r-1} \Gamma_{(\mathrm{r}-1,\mathrm{k})}(\| u_{0k} \|_\infty).
\end{aligned}$$

These bounds ensure that a solution to (24) exists for all time and space. □

Note that for $r = 2$, the estimates in (29) reduce to the previous results in (20) by Theorem 3.2:

$$\begin{aligned}
\| u_1 \|_\infty &\leq \| u_{01} \|_\infty + z_1(\| u_{02} \|_\infty), \\
\| u_2 \|_\infty &\leq \| u_{02} \|_\infty + \gamma_1(\| u_{01} \|_\infty).
\end{aligned}$$

5. Initial–boundary value problem. We define the initial-boundary value problem for the $r \times r$ system to be the following:

$$
\begin{aligned}
(\partial_t + \lambda_1 \partial_x) u_1(x,t) &= f_1(u_1, u_2), \\
&\vdots \qquad\qquad \vdots \\
(\partial_t + \lambda_i \partial_x) u_i(x,t) &= -f_{i-1}(u_{i-1}, u_i) + f_i(u_i, u_{i+1}), \\
&\vdots \qquad\qquad \vdots \\
(\partial_t + \lambda_r \partial_x) u_r(x,t) &= -f_{r-1}(u_{r-1}, u_r), \\
\vec{u}(x,0) &= \vec{u}_0(x),
\end{aligned} \tag{30}
$$

where λ_i is arbitrary and $x \in [0, +\infty)$. For those $\lambda_i > 0$, we prescribe the following boundary data:

$$u_i(0,t) = h_i(t). \tag{31}$$

PROPOSITION 5.1 (local existence). *Consider the initial-boundary value problem* (30), (31). *For* $k = 1, \ldots, r-1$, *let each* $f_k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *be* C^∞, *and let the components of the initial data,* $\vec{u}_0(x)$, *and boundary data,* $\vec{h}(t)$, *be uniformly bounded, continuously differentiable functions with uniformly bounded derivatives. If for those* $\lambda_i > 0$ *the data are compatible at the origin, that is, if*

$$\lim_{s \to 0^+} h_i(s) = \lim_{s \to 0^+} u_{0i}(s)$$

and

$$\lim_{s \to 0^+} h_i'(s) = \lim_{s \to 0^+} [-\lambda_i u_{0i}(s) + F_i(u_i(s,0), u_{i+1}(s,0))],$$

where F_i *is the source term for the* i*th equation in* (30), *then there exists a* $\delta > 0$ *such that the initial-boundary value problem* (30), (31) *has a unique solution for all* $x \geq 0$ *and* $t \in [0, \delta)$. *If in addition the initial data are nonnegative,* $C^1_\infty(\mathbb{R}^+)$ *functions and each* f_k *is a reaction function, then the solution to the initial-boundary value problem is a nonnegative,* $C^1_\infty([0,\delta) \times \mathbb{R}^+)$ *function.*

Proof. See [5] and [8]. □

THEOREM 5.2 (global existence). *Consider the initial-boundary value problem* (30), (31) *and assume the hypotheses defined in Proposition* 5.1. *If* $\vec{F} \equiv (F_1, \ldots, F_r)$ *satisfies the stability criteria, then a solution to* (30), (31) *exists for all* $x \geq 0$ *and for all* $t \geq 0$.

Proof. To prove this theorem, we apply the same argument used in Theorem 4.2, where the energy norms are modified to account for the boundary data at the origin.

For some $T > 0$, assume that a nonnegative solution exists to (30), (31) for all $x \geq 0$ and $t \in [0, T]$.

For those $\lambda_i \leq 0$ where $i = 1, \ldots, r$, we modify the energy norms used in Theorem 4.2 in the following way:

$$
\begin{aligned}
E_{u_i}(t) \equiv \int_0^\infty \Bigg(u_i^n(x,t) + \sum_{k=1}^{i-1} \int_0^{u_k(x,t)} & n\Gamma^{n-1}_{(\mathrm{i-1,k})}(\tau) d\tau \\
& + \sum_{k=i+1}^{r} \int_0^{u_k(x,t)} n\mathcal{Z}^{n-1}_{(\mathrm{i,k-1})}(\tau) d\tau \Bigg) dx
\end{aligned}
$$

$$+\sum_{\substack{k=1\\ \lambda_k>0}}^{i-1}\int_t^T\left[\int_0^{h_k(s)} n\lambda_k\Gamma^{n-1}_{(i-1,k)}(\tau)d\tau\right]ds$$
$$+\sum_{\substack{k=i+1\\ \lambda_k>0}}^{r}\int_t^T\left[\int_0^{h_k(s)} n\lambda_k\mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau\right]ds,$$

where $n>1$. Note that for $i=1$, those sums whose limits go from $k=1$ to $k=0$ are set to zero. Similarly, for $i=r$, we set those sums from $k=r+1$ to r to be zero. Differentiating E_{u_i} with respect to time yields

$$\frac{d}{dt}E_{u_i}(t)=\int_0^\infty\left(nu_i^{n-1}(-f_{i-1}+f_i)+n\Gamma^{n-1}_{(i-1,1)}(u_1)f_1\right)dx$$
$$+\int_0^\infty\left(\sum_{k=2}^{i-1}n\Gamma^{n-1}_{(i-1,k)}(u_k)(-f_{k-1}-f_k)\right.$$
$$\left.+\sum_{k=i+1}^{r-1}n\mathcal{Z}^{n-1}_{(i,k-1)}(u_k)(-f_{k-1}+f_k)\right)dx \tag{32}$$
$$+\int_0^\infty\left(n\mathcal{Z}^{n-1}_{(i,r-1)}(u_r)(-f_{r-1})\right)dx$$
$$+\int_0^\infty\left(-\lambda_i\partial_x[u_i^n]-\sum_{k=1}^{i-1}\partial_x\left[\int_0^{u_k(x,t)}n\lambda_k\Gamma^{n-1}_{(i-1,k)}(\tau)d\tau\right]\right.$$
$$\left.-\sum_{k=i+1}^{r}\partial_x\left[\int_0^{u_k(x,t)}n\lambda_k\mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau\right]\right)dx \tag{33}$$
$$-\sum_{\substack{k=1\\ \lambda_k>0}}^{i-1}\int_0^{h_k(t)}n\lambda_k\Gamma^{n-1}_{(i-1,k)}(\tau)d\tau-\sum_{\substack{k=i+1\\ \lambda_k>0}}^{r}\int_0^{h_k(t)}n\lambda_k\mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau.$$

Collecting boundary terms (33), we find as in (27) of Theorem 4.2 that the boundary terms at $+\infty$ vanish since the solution is in C^1_∞. However, the boundary terms at the origin do not vanish and we are left with

$$\lambda_i u_i^n(0,t)+\sum_{\substack{k=1\\ \lambda_k\le 0}}^{i-1}\int_0^{u_k(0,t)}n\lambda_k\Gamma^{n-1}_{(i-1,k)}(\tau)d\tau+\sum_{\substack{k=1\\ \lambda_k>0}}^{i-1}\int_0^{u_k(0,t)}n\lambda_k\Gamma^{n-1}_{(i-1,k)}(\tau)d\tau$$
$$+\sum_{\substack{k=i+1\\ \lambda_k\le 0}}^{r}\int_0^{u_k(0,t)}n\lambda_k\mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau+\sum_{\substack{k=i+1\\ \lambda_k>0}}^{r}\int_0^{u_k(0,t)}n\lambda_k\mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau$$
$$-\sum_{\substack{k=1\\ \lambda_k>0}}^{i-1}\int_0^{h_k(t)}n\lambda_k\Gamma^{n-1}_{(i-1,k)}(\tau)d\tau-\sum_{\substack{k=i+1\\ \lambda_k>0}}^{r}\int_0^{h_k(t)}n\lambda_k\mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau.$$

For all $\lambda_k>0$, the solution $u_k(0,t)$ is the prescribed boundary data $h_k(t)$. So the

boundary terms further reduce to

$$\lambda_i u_i^n(0,t) + \sum_{\substack{k=1\\ \lambda_k \le 0}}^{i-1} \int_0^{u_k(0,t)} n\lambda_k \Gamma^{n-1}_{(i-1,k)}(\tau)d\tau + \sum_{\substack{k=i+1\\ \lambda_k \le 0}}^{r} \int_0^{u_k(0,t)} n\lambda_k \mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau. \tag{34}$$

By assumption $\lambda_i \le 0$ and the sums are over those $\lambda_k \le 0$, so the terms in (34) are nonpositive. Thus, the energy norms were modified so that the positive growth factors at the origin vanish and the remaining terms either do not contribute to or decrease the growth of the energy norm as time progresses.

By applying the same manipulations used in Theorem 4.2 to the sums in (32) and combining (34), we can easily show that $E_{u_i}(t)$ is nonincreasing in time. Using the same argument in Theorem 3.2 yields the following a priori bound for solution $u_i(x,t)$, where $\lambda_i \le 0$:

$$\begin{aligned} \| u_i(x,t) \|_\infty \le \| u_{0i} \|_\infty + \sum_{k=1}^{i-1} \Gamma_{(i-1,k)}(\| u_{0k} \|_\infty) + \sum_{k=i+1}^{r} \mathcal{Z}_{(i,k-1)}(\| u_{0k} \|_\infty) \\ + \sum_{\substack{k=1\\ \lambda_k > 0}}^{i-1} \Gamma_{(i-1,k)}(\max_{0\le s\le T} h_k(s)) + \sum_{\substack{k=i+1\\ \lambda_k > 0}}^{r} \mathcal{Z}_{(i,k-1)}(\max_{0\le s\le T} h_k(s)). \end{aligned} \tag{35}$$

For those $\lambda_i > 0$ where $i = 1, \dots, r$, we have the following energy norms:

$$\begin{aligned} E_{u_i}(t) \equiv \int_0^\infty \Bigg(u_i^n(x,t) + \sum_{k=1}^{i-1} \int_0^{u_k(x,t)} n\Gamma^{n-1}_{(\mathrm{i-1,k})}(\tau)d\tau \\ + \sum_{k=i+1}^{r} \int_0^{u_k(x,t)} n\mathcal{Z}^{n-1}_{(\mathrm{i,k-1})}(\tau)d\tau \Bigg) dx \\ + \int_t^T \lambda_i h_i^n(s)ds + \sum_{\substack{k=1\\ \lambda_k > 0}}^{i-1} \int_t^T \left[\int_0^{h_k(s)} n\lambda_k \Gamma^{n-1}_{(i-1,k)}(\tau)d\tau \right] ds \\ + \sum_{\substack{k=i+1\\ \lambda_k > 0}}^{r} \int_t^T \left[\int_0^{h_k(s)} n\lambda_k \mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau \right] ds, \end{aligned}$$

where $n > 1$. Again, for $i = 1$ and $i = r$, we set the appropriate sums equal to zero. By differentiating E_{u_i} with respect to time and collecting boundary data, we have

$$\begin{aligned} \lambda_i u_i^n(0,t) + \sum_{\substack{k=1\\ \lambda_k \le 0}}^{i-1} \int_0^{u_k(0,t)} n\lambda_k \Gamma^{n-1}_{(i-1,k)}(\tau)d\tau + \sum_{\substack{k=1\\ \lambda_k > 0}}^{i-1} \int_0^{u_k(0,t)} n\lambda_k \Gamma^{n-1}_{(i-1,k)}(\tau)d\tau \\ + \sum_{\substack{k=i+1\\ \lambda_k \le 0}}^{r} \int_0^{u_k(0,t)} n\lambda_k \mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau + \sum_{\substack{k=i+1\\ \lambda_k > 0}}^{r} \int_0^{u_k(0,t)} n\lambda_k \mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau \\ - \lambda_i h_i^n(t) - \sum_{\substack{k=1\\ \lambda_k > 0}}^{i-1} \int_0^{h_k(t)} n\lambda_k \Gamma^{n-1}_{(i-1,k)}(\tau)d\tau - \sum_{\substack{k=i+1\\ \lambda_k > 0}}^{r} \int_0^{h_k(t)} n\lambda_k \mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau. \end{aligned}$$

Again, all the positive growth factors at the origin vanish, and we are left with

$$\sum_{\substack{k=1\\ \lambda_k \le 0}}^{i-1} \int_0^{u_k(0,t)} n\lambda_k \Gamma^{n-1}_{(i-1,k)}(\tau)d\tau + \sum_{\substack{k=i+1\\ \lambda_k \le 0}}^{r} \int_0^{u_k(0,t)} n\lambda_k \mathcal{Z}^{n-1}_{(i,k-1)}(\tau)d\tau,$$

which consists of nonpositive terms. The a priori bound for solution $u_i(x,t)$, where $\lambda_i > 0$, is

$$\begin{aligned}(36)\quad \| u_i(x,t) \|_\infty \le \| u_{0i} \|_\infty + \max_{0\le s\le T} h_i(s) + \sum_{k=1}^{i-1} \Gamma_{(i-1,k)}(\| u_{0k} \|_\infty) \\ + \sum_{k=i+1}^{r} \mathcal{Z}_{(i,k-1)}(\| u_{0k} \|_\infty) + \sum_{\substack{k=1\\ \lambda_k>0}}^{i-1} \Gamma_{(i-1,k)}(\max_{0\le s\le T} h_k(s)) \\ + \sum_{\substack{k=i+1\\ \lambda_k>0}}^{r} \mathcal{Z}_{(i,k-1)}(\max_{0\le s\le T} h_k(s)).\end{aligned}$$

Global existence of a solution to (30) follows easily. □

Note that the a priori estimates (35), (36) depend on the initial data, boundary data, and on the geometry of the zero set of the source terms. As a simple example, consider the 2×2 system:

$$\begin{aligned}(\partial_t + \lambda_1\partial_x)u_1(x,t) &= f(u_1,u_2)\\ (\partial_t + \lambda_2\partial_x)u_2(x,t) &= -f(u_1,u_2)\\ \vec{u}(x,0) &= \vec{u}_0(x)\\ u_1(0,t) &= h_1(t), \qquad t>0,\end{aligned}$$

where $\lambda_1 > 0$, $\lambda_2 \le 0$, and the initial and boundary data are compatible at the origin as defined in Proposition 5.1. The energy norms as defined in Theorem 5.2 reduce to

$$E_{u_1}(t) \equiv \int_0^\infty \left(u_1^n(x,t) + \int_0^{u_2(x,t)} nz^{n-1}(\tau)d\tau \right) dx + \int_t^T \lambda_1 h_1^n(s)ds,$$

$$E_{u_2}(t) \equiv \int_0^\infty \left(u_2^n(x,t) + \int_0^{u_1(x,t)} n\gamma^{n-1}(\tau)d\tau \right) dx + \int_t^T \int_0^{h_1(s)} n\lambda_1\gamma^{n-1}(\tau)d\tau ds$$

and the a priori bounds given by (35), (36) are

$$\| u_1(x,t) \|_\infty \le \| u_{01}(x) \|_\infty + \max_{0\le s\le T} h_1(s) + z(\| u_{02} \|_\infty),$$

$$\| u_2(x,t) \|_\infty \le \| u_{02}(x) \|_\infty + \gamma(\| u_{01} \|_\infty) + \gamma(\max_{0\le s\le T} h_1(s)).$$

Acknowledgment. The author wishes to thank Michael C. Reed for suggesting this problem, much of which has appeared in the author's dissertation.

REFERENCES

[1] J. Blum, D. Carr, and M. Reed, *Theoretical analysis of lipid transport in sciatic nerve*, Biochimica et Biophysica Acta, 1125 (1992), pp. 313–320.

[2] J. BLUM AND M. REED, *A model for fast axonal transport*, J. Cell Motility, 5 (1985), pp. 507–527.
[3] ———, *The transport of organelles in axons*, J. Math. Biosciences, 90 (1988), pp. 233–245.
[4] H. CABANNES, *Solution globale du problème de Cauchy en théorie cinétique discrète*, Journal de Mécanique, 17 (1978), pp. 1–22.
[5] D. CARR, *Reaction–Hyperbolic Systems in One Space Dimension*, Ph.D. thesis, Department of Mathematics, Duke University, Durham, NC, 1992.
[6] R. ILLNER, *Global existence for two-velocity models of the Boltzmann equation*, Math. Methods Appl. Sci., 1 (1979), pp. 187–193.
[7] A. LEHNINGER, *Biochemistry*, 2nd ed., Worth Publishers, Inc., New York, 1975.
[8] J. RAUCH AND F. MASSEY, *Differentiability of solutions to hyperbolic initial-boundary value problems*, Trans. Amer. Math. Soc., 189 (1974), pp. 303–318.
[9] M. REED, 1990, private communication.
[10] M. REED AND J. BLUM, *Theoretical analysis of radioactivity profiles during fast axonal transport: Effects of deposition and turnover*, Cell Motility and the Cytoskeleton, 6 (1986), pp. 620–627.
[11] M. REED, S. VENAKIDES, AND J. BLUM, *Approximate traveling waves in linear reaction-hyperbolic equations*, SIAM J. Appl. Math., 50 (1990), pp. 167–180.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 415–427, March 1995

008

GLOBAL ATTRACTORS FOR PARABOLIC PROBLEMS IN FRACTIONAL POWER SPACES*

ALEXANDRE NOLASCO DE CARVALHO† AND JOSÉ GASPAR RUAS-FILHO†

Abstract. This paper deals with global well-posedness and existence of global attractors for systems of weakly coupled semilinear parabolic problems in fractional power spaces which are embedded in L^∞. In these spaces, no growth assumption on the nonlinearity is required for local existence and it can be proven that some sort of dissipation takes place. The tools employed are the theory of invariant regions and the invariance theory. The first provides global existence of solutions whereas the second provides point dissipativeness. Some applications to chemical kinetic problems are considered, as are some problems arising as limiting problems for reaction-diffusion equations in thin domains around a point.

Key words. global attractors, cooperative systems, thin domains, fractional power spaces, reaction-diffusion equations

AMS subject classifications. 35B40, 34C35, 35K57, 58B39

1. Introduction and statement of the results.

Let Ω be a bounded smooth domain of $\mathbb{R}^n$, $n \leq 3$. In this paper we consider parabolic problems of the form

$$\begin{cases} u_t = d\Delta u - \gamma u + f(u) & \text{in} \quad \Omega, \\ \dfrac{\partial u}{\partial n} = 0 & \text{in} \quad \partial\Omega, \end{cases} \tag{1.1}$$

where d and γ are positive constants.

We are interested in global well-posedness and the existence of global attractors for such problems. To simplify the presentation we assume that u is scalar and the nonlinearity $f : \mathbb{R} \to \mathbb{R}$ is a C^1 function that satisfies

$$\limsup_{|u|\to\infty} \frac{f(u)}{u} \leq -\delta < 0. \tag{1.2}$$

This dissipativeness condition will play a fundamental role in proving that there is an absorbing set for (1.1).

To describe the results we introduce some terminology. Let $X = L^2(\Omega)$ and $A : D(A) \subset X \to X$ be the self-adjoint operator defined by

$$D(A) = \{\phi \in H^2(\Omega) : \tfrac{\partial \phi}{\partial n} = 0 \text{ in } \partial\Omega\},$$

$$A\phi = -d\Delta\phi + \gamma\phi \quad \forall \phi \in D(A).$$

This operator is sectorial, and we can define its fractional powers A^α, $0 \leq \alpha$, and the associated fractional power spaces $X^\alpha = D(A^\alpha)$ endowed with the graph norm. We will be concerned only with $\alpha < 1$.

* Received by the editors March 29, 1993; accepted for publication (in revised form) September 23, 1993.

† Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, 13560-970, São Carlos, SP, Brazil. The research of the first author was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil, under grant 300889/92-5.

It is well known that under some growth assumptions on the nonlinearity f, the problem (1.1) has a global attractor in $H^1(\Omega)$. More specifically, if f satisfies

$$|f(u) - f(v)| \leq c(e^{\theta|u|^\eta} + e^{\theta|v|^\eta})|u - v|, \quad \eta < 2 \text{ and } \theta > 0, \text{ if } n = 2,$$

$$|f(u) - f(v)| \leq c(1 + |u|^2 + |v|^2)|u - v| \quad \text{if } n = 3,$$

then the problem (1.1) has a global attractor in $H^1(\Omega)$. (See, for example, Hale [14], Hale and Raugel [15], Carvalho [3], and Carvalho and Oliveira [5].)

These growth assumptions are necessary to obtain local existence of solutions for (1.1) and also play a role in obtaining some energy estimates necessary to guarantee that the solution operator for (1.1) defines a global dynamical system which is bounded dissipative.

It would be interesting to pose the problem (1.1) in a space where no growth assumption on the nonlinearity f was required for local existence and where we were able to prove the existence of a global attractor.

We will consider spaces X^α which are embedded in $L^\infty(\Omega)$. Our aim is to show that in such spaces the problem (1.1) has a global attractor and to obtain some good estimates for the size of the attractor in the uniform norm.

Hale [13] proved the existence of a local attractor for (1.1), which coincides with the embedding of the attractor for $\dot{u} = f(u)$ into the subspace of constant functions of X^α, $\alpha > \frac{3}{4}$, if the diffusion coefficient d is large (see also Hale and Rocha [16], [17] and Hale and Sakamoto [18]). However, the techniques employed by Hale [13] would only apply to global attractors if some a priori bound on the size of the absorbing set could be obtained and only if the diffusion coefficient is large (see Carvalho [3] and Carvalho and Oliveira [5]).

We prove the existence of a global attractor for the problem (1.1) regardless of the size of d. We also give uniform (with respect to d) bounds (in L^∞) on the size of the attractor which will make the results of Carvalho [3] and Carvalho and Oliveira [5] applicable to the case $\alpha \neq \frac{1}{2}$ as in Hale [13], Hale and Rocha [16], [17], and Hale and Sakamoto [18] (see also Fusco [11]).

To carry on this project, we need to obtain that the solution operator associated to (1.1) is globally defined, that orbits of bounded subsets of X^α under the flow defined by (1.1) are bounded subsets of X^α, and that there is a bounded set that attracts points of X^α. Since the solution operator associated to (1.1) is compact, Theorem 3.4.6 in Hale [14] would guarantee the existence of a global attractor.

Recall (see Hale [14], for example) that if $T(t) : X \to X$, $t \geq 0$, is a semigroup of transformations on a Banach space X, then a set $\mathcal{A}$ is an attractor if it is a compact invariant set that attracts a neighborhood $\mathcal{O}$ of itself; that is, $\mathcal{A}$ is compact, $T(t)\mathcal{A} = \mathcal{A}$ for $t \geq 0$, and there is a neighborhood $\mathcal{O}$ of $\mathcal{A}$ such that dist $(T(t)U, \mathcal{A}) \to 0$ as $t \to \infty$. The set $\mathcal{A}$ is a global attractor if it attracts each bounded set of X.

Next, we state our main results. Consider the following system of reaction diffusion equations with dispersion:

$$(1.1)' \qquad \begin{cases} u_t = D\Delta u - \gamma u + \sum_{j=1}^n B_j(x)\frac{\partial u}{\partial x_j} + f(u) & \text{in } \Omega, \\ \\ \dfrac{\partial u}{\partial n} = 0 & \text{in } \partial\Omega, \end{cases}$$

where $u = (u_1, u_2, \ldots, u_N)^\top$, $N \geq 1$, $D = \operatorname{diag}(d_1, \ldots, d_N)$, $d_i > 0$, $1 \leq i \leq N$, and $B_j = \operatorname{diag}(b_j^1, \ldots, b_j^N)$ is continuous in $\bar{\Omega}$, $1 \leq j \leq n$. The nonlinearity $f =$

$(f_1, \dots, f_N)^\top : \mathbb{R}^N \to \mathbb{R}^N$ is assumed to be a C^1 function that satisfies

$$\limsup_{|u_i| \to \infty} \frac{f_i(u)}{u_i} \leq -\delta < 0 \tag{1.3}$$

uniformly with respect to $\hat{u}_i = (u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N) \in \mathbb{R}^{N-1}$, $1 \leq i \leq N$. In §2, for suitable values of γ, we prove the following result.

THEOREM 1.1. *Under the above hypotheses, the solutions of* (1.1)′ *with initial data in* X^α *are globally defined and orbits of bounded subsets of* X^α *under the flow defined by* (1.1)′ *are also bounded subsets of* X^α.

In §3 we show the existence of a global attractor neglecting the dispersion terms in (1.1)′. Indeed, we prove the following theorem.

THEOREM 1.2. *Suppose that in* (1.1)′, $B_j \equiv 0$, $1 \leq j \leq n$. *Then the solution operator* $\{T(t), t \geq 0\}$ *for* (1.1)′ *is point dissipative and therefore it has a global attractor.*

In §4 we consider the structure of gradient systems that such problems have and prove the next theorem.

THEOREM 1.3. *Let* $\Sigma_i = [-\bar{\xi}_i, \xi_i]$, *where* ξ_i, $\bar{\xi}_i$ *are positive constants,* $1 \leq i \leq N$. *If* $s_i f_i(u_1, \dots, u_{i-1}, s, u_{i+1}, \dots, u_N) < 0$ *whenever* $s_i \notin \Sigma_i$, *then* $\phi(x) \in \Sigma := \Sigma_1 \times \cdots \times \Sigma_N$, $\forall x \in \Omega$, *and* $\forall \phi \in \mathcal{A}$.

Finally, in §5 we consider applications to a class of cooperative systems arising from thin domain problems around a point.

The tools employed are the invariance theory as in Henry [19] and the theory of invariant regions of Chueh, Conley, and Smoller [7].

We believe that the hypothesis (1.2) can be relaxed a little to include the case when the nonlinearity f satisfies $s f(s) - \gamma s^2 < 0$, $|s| \geq \xi$, for some $\xi > 0$. However, we have not been able to prove that, in this case, orbits of bounded subsets of X^α under the flow defined by (1.1) are bounded subsets of X^α.

We remark that other boundary conditions can be considered with little change; hence, we will restrict the presentation to the homogeneous Neumann boundary conditions case. We also remark that if the nonlinearity depends on the spatial variable, the results hold with almost no change in the proofs.

The following result identifies the values of α that we will be considering; its proof can be found in Henry [19, Thm. 1.6.1].

THEOREM 1.4. *Suppose that* $\Omega \subset \mathbb{R}^n$ *is an open bounded set with smooth boundary. Then for* $0 \leq \alpha \leq 1$,

$$X^\alpha \subset W^{1,2}(\Omega) \quad \textit{when} \quad \frac{1}{2} \leq \alpha,$$

$$X^\alpha \subset C^\nu(\Omega) \quad \textit{when} \quad 0 \leq \nu + \frac{n}{2} < 2\alpha.$$

Furthermore, the embedding is compact whenever the inequality is strict.

Therefore, if we assume that $\alpha > \max\{\frac{n}{4}, \frac{1}{2}\}$ for $n = 2, 3$, or $\alpha \geq \frac{1}{2}$ for $n = 1$, we have that

$$X^\alpha \subset H^1(\Omega) \cap L^\infty(\Omega). \tag{1.4}$$

Hereafter we assume that (1.4) holds. The next lemma is the main reason why we are interested in working with such spaces.

LEMMA 1.5. *Let $f : \mathbb{R} \to \mathbb{R}$ be a C^1 function and $f^e : X^\alpha \to X$ be the map defined by*

$$f^e(\phi)(x) = f(\phi(x)).$$

Then, f^e is a well-defined compact map which is Lipschitz continuous in bounded sets of X^α. Furthermore, for any $r > 0$ there exists a constant N_1, depending only on r, such that

$$\|f^e(\phi)\|_{L^\infty(\Omega)} \leq N_1,$$

whenever $\|\phi\|_X^\alpha \leq r$.

The proof of this result is rather trivial and we omit it. This lemma states that the problem (1.1) is locally well posed in X^α even if no growth assumption is made on the nonlinearity f.

2. Proof of Theorem 1.1. In this section we prove that solutions of (1.1)′ with initial data in X^α are globally defined and that orbits of bounded subsets of X^α under the flow defined by (1.1)′ are also bounded subsets of X^α. To prove this result, we will use the following lemma.

LEMMA 2.1. *Let A be a sectorial operator and $f^e : X^\alpha \to X$ be a bounded map which is Lipschitz continuous in bounded subsets of X^α. Then, the problem*

$$\begin{aligned} &\dot{u} + Au = f^e(u), \\ &u(0) = u_0 \in X^\alpha, \end{aligned} \tag{2.1}$$

has a local solution $T(t)u_0$ defined in a maximal interval of existence $[0, t_{\max})$. Furthermore, either $\|T(t)u_0\|_\alpha \to \infty$ when $t \to \infty$ or $t_{\max} = +\infty$.

For a proof of this lemma see Henry [19]. We observe that it is not enough that the nonlinearity f^e be locally Lipschitz continuous. We assume that f is Lipschitz continuous in bounded sets, which is suitable for our applications. Some extra hypothesis is necessary because the phase space is not locally compact.

We know that $A = \operatorname{diag}(A_1, \ldots, A_N)$ defined by

$$D(A_i) = \{\phi \in H^2(\Omega) : \tfrac{\partial \phi}{\partial n} = 0\},$$

$$-A_i\phi = d_i\Delta\phi - \gamma\phi + \textstyle\sum_{j=1}^n b_j^i(x)\frac{\partial \phi}{\partial x_j},$$

generates an analytic semigroup on X^α and that it satisfies the following estimates:

$$\begin{aligned} &\|e^{-At}u_0\|_X^\alpha \leq Me^{-\epsilon t}\|u_0\|_X^\alpha, \quad t \geq 0, \\ &\|e^{-At}u_0\|_X^\alpha \leq Me^{-\epsilon t}t^{-\alpha}\|u_0\|_X, \quad t > 0 \end{aligned} \tag{2.2}$$

for some $\epsilon > 0$, $M \geq 1$.

By writing the problem (1.1)′ in the form (2.1) and using the variation of constants formula, we can view its solution through $u_0 \in X^\alpha$ as

$$T(t)u_0 = e^{-At}u_0 + \int_0^t e^{-A(t-s)}f^e(T(s)u_0)ds,$$

where $(f^e(\phi))(x) = f(\phi(x))$ for all $\phi \in X^\alpha$. If $\|T(t)u_0\|_{L^\infty(\Omega)} \leq N_1$, $t \in [0, t_{\max})$ for some $N_1 > 0$, we have that

$$\|T(t)u_0\|_X^\alpha \leq M\|u_0\|_\alpha e^{-\epsilon t} + MK\int_0^t e^{-\epsilon(t-s)}(t-s)^{-\alpha}ds, \tag{2.3}$$

where $K = \sup_{t\in[0,t_{\max})} \|f^e(T(t)u_0)\|_{L^2(\Omega)}$. Therefore, if we are able to obtain estimates in $L^\infty(\Omega)$ for $T(t)u_0$, $0 \le t < t_{\max}$, a similar estimate can be obtained in X^α and the solutions are globally defined.

To obtain such estimates in $L^\infty(\Omega)$ we introduce the notion of invariant regions as in Smoller [24].

DEFINITION 2.2. *A set $\Sigma \subset \mathbb{R}^N$ is called a positively invariant region for the local solution of (1.1)′ if any solution $T(t)u_0$ that satisfies $u_0(x) \in \Sigma$, $\forall x \in \Omega$ is such that $(T(t)u_0)(x) \in \Sigma$, $\forall x \in \Omega$ and for all t in the maximal interval of existence of the solution.*

Our next result characterizes some of the invariant regions of the problem (1.1)′. Its proof follows Smoller [24] and is presented here for the sake of completeness.

THEOREM 2.3. *Let $\bar\xi_j$, $\xi_j > 0$, $1 \le j \le N$ be such that $u_j f_j(u) < 0$ for all $u \in \mathbb{R}^N$ with $u_j \notin [-\bar\xi_j, \xi_j]$. Then the rectangle $\Sigma = [-\bar\xi_1, \xi_1] \times [-\bar\xi_2, \xi_2] \times \cdots \times [-\bar\xi_N, \xi_N]$ is an invariant region for the local solution of (1.1)′.*

Proof. If there is a solution $v(x,t) = (v^1(x,t), v^2(x,t), \dots, v^N(x,t))$ of (1.1)′ with initial data $v(x,0) = (v^1(x,0), v^2(x,0), \dots, v^N(x,0)) \in \Sigma$ for all $x \in \bar\Omega$, that does not stay in $\Sigma_1 = (-\infty, \xi_1] \times \mathbb{R}^{N-1}$ for all $t \in [0, t_{\max})$, then there is a t_0 and $x_0 \in \Omega$ such that

$$v^1(x,t) < \xi_1, \quad 0 \le t < t_0, \quad x \in \Omega, \quad \text{and} \quad v^1(x_0,t_0) = \xi_1.$$

We observe that x_0 need not to be in Ω. If this is the case, by a change of variables we can assume that $\frac{\partial u}{\partial n} = -\zeta < 0$. Then the maximum happens in Ω and the theorem will follow if we let $\zeta \to 0$.

Therefore, if $v^1(x_0,t) < \xi_1, \forall t \in [0,t_0)$ and $v^1(x_0,t_0) = \xi_1$ implies that $v^1_t(x_0,t_0) < 0$, then Σ_1 is invariant.

Consider the following:

$$v^1_t(x_0,t_0) = d_1 \Delta v^1(x_0,t_0) - \gamma v^1(x_0,t_0) + \sum_{j=1}^{n} B^1_j(x_0) \frac{\partial v^1}{\partial x_j}(x_0,t_0) + f_1(v(x_0,t_0)). \tag{2.4}$$

We claim that $\nabla v^1 = 0$ at (x_0,t_0). In fact, if $\frac{\partial v^1}{\partial x_i} > 0$ at (x_0,t_0) for some $1 \le i \le n$, then $v^1(x_0,t_0) = \xi_1$ and $v^1(x,t_0) > \xi_1$ for some x with $|x - x_0|$ small. This implies that $v^1(x,t) > \xi_1$ for x near x_0 and $t < t_0$ near t. This contradicts the definition of t_0 and $\frac{\partial v^1}{\partial x_i} \le 0$. By using the same reasoning, we obtain that $\frac{\partial v^1}{\partial x_i} < 0$ at (x_0,t_0) for some $1 \le i \le n$ leads to a contradiction and the claim is proved.

Similarly, $v^1_{x_i x_i} \le 0$ for all $1 \le i \le n$. Therefore, $\Delta v^1(x_0,t_0) \le 0$. From expression (2.4), we have that

$$v^1_t(x_0,t_0) \le f_1(v(x_0,t_0)) - \gamma v^1(x_0,t_0).$$

Since $f_1(v(x_0,t_0)) - \gamma v^1(x_0,t_0) < 0$, we conclude that $v^1_t(x_0,t_0) < 0$.

From the reasoning at the beginning of the proof we have that $(-\infty, \xi_1] \times \mathbb{R}^{N-1}$ is an invariant region for the local solution of (1.1)′. In the same way, we obtain that $[-\xi_1, \infty) \times \mathbb{R}^{N-1}$ is also an invariant region. From the fact that the intersection of invariant regions is still invariant, the proof is completed. □

This theorem shows that for any $u_0 \in X^\alpha$ the local solution $T(t)u_0$ of (1.1)′ through u_0 satisfies

$$\|T(t)u_0\|_{L^\infty(\Omega)} \le N_1$$

for some $N_1 > 0$, whenever defined.

It follows from both Lemma 2.1 and expression (2.4) that the problem (1.1)′ defines a global dynamical system in X^α.

Since the semigroup generated by A satisfies (2.2) for some $\epsilon > 0$, our following computations show that orbits of bounded subsets of X^α are bounded subsets of X^α. To do this, we resort once more to the variation of constants formula and to Lemma 2.1.

Let B be a bounded subset of X^α. From the fact that X^α is embedded in $L^\infty(\Omega)$ and from the variation of constant formula, we have that there are constants $K_i > 0$, $1 \leq i \leq 4$ depending only on B such that

$$\begin{aligned}\|T(t)u_0\|_X^\alpha &\leq K_1 + K_2 \int_0^t (t-s)^{-\alpha} e^{-\epsilon(t-s)} \|f(T(s)u_0)\|_{L^\infty(\Omega)} ds \\ &\leq K_1 + K_3 \int_0^t (t-s)^{-\alpha} e^{-\epsilon(t-s)} ds\end{aligned}$$

$\forall t \geq 0$ and $\forall u_0 \in B$. We used that $\|T(t)u_0\|_{L^\infty(\Omega)} \leq K_4$, which follows from the embedding of X^α into $L^\infty(\Omega)$ and from Theorem 2.3. This proves Theorem 1.1.

In what follows, we consider an example of a reaction-diffusion equation with dispersion for which we know that orbits of bounded sets are bounded.

Consider a system of reaction-diffusion equations of chemical kinetics (see, for example, Chueh, Conley, and Smoller [7])

$$\begin{aligned} w_t &= d_1 \Delta w + \textstyle\sum_{i=1}^n \beta_i^1 \frac{\partial w}{\partial x_i} + g(w,v) \qquad \text{in} \quad \Omega, \\ v_t &= d_1 \Delta v + \textstyle\sum_{i=1}^n \beta_i^2 \frac{\partial v}{\partial x_i} + h(w,v) \qquad \text{in} \quad \Omega, \\ \frac{\partial w}{\partial n} &= \frac{\partial v}{\partial n} = 0 \qquad \text{in} \quad \partial\Omega, \end{aligned} \tag{2.5}$$

where $\Omega \subset \mathbb{R}^n$, $n \leq 3$ is as in §1, and $f, g : \mathbb{R}^2 \to \mathbb{R}$ are C^1-functions that satisfy

$$\limsup_{|u|\to\infty} \frac{g(u,v)}{u} = -\infty, \qquad \limsup_{|v|\to\infty} \frac{h(u,v)}{v} = -\infty,$$

where the first limit is uniform with respect to v and the second is uniform with respect to w.

Then the system (2.5) can be rewritten as

$$\begin{aligned} u_t &= D\Delta u - \gamma u + \textstyle\sum_{j=1}^n B_j \frac{\partial u}{\partial x_j} + f(u) \quad \text{in} \quad \Omega, \\ \frac{\partial u}{\partial n} &= 0 \quad \text{in} \quad \partial\Omega, \end{aligned}$$

where

$$u = (w,v)^\top \in \mathbb{R}^2, \quad B_j = \operatorname{diag}(\beta_j^1, \beta_j^2), \quad f(u) = \begin{pmatrix} g(w) + \gamma w \\ h(v) + \gamma v \end{pmatrix} : \mathbb{R}^2 \to \mathbb{R}^2$$

satisfy (1.3) and γ is chosen as follows.

Consider the operator A defined by $A = \operatorname{diag}(A_1, A_2)$,

$$D(A_j) = \{\phi \in H^2(\Omega) : \tfrac{\partial \phi}{\partial n} = 0\},$$

$$A_j \phi = d_j \Delta \phi - \gamma \phi + \textstyle\sum_{i=1}^n \beta_i^j \frac{\partial \phi}{\partial x_i}, \quad j = 1, 2.$$

Let γ be such that the analytic semigroup generated by A decays exponentially to zero as $t \to \infty$. The results in this section imply that the solution operator $S(t)$ for (2.5) is defined globally and orbits of bounded sets under $\{S(t) : t \geq 0\}$ are bounded subsets of X^α.

3. Proof of Theorem 1.2. In this section we use the invariance theory as in Henry [19] to prove that there is a bounded set in X^α which attracts points of X^α. Unfortunately, the techniques employed in this section will not work for systems of reaction-diffusion equations with dispersion due to the fact that we will not be able to find a Lyapunov function for such systems. We state the results of the invariance theory that we will use by starting with the definition of Lyapunov function.

DEFINITION 3.1. *Let $\{S(t),\ t \geq 0\}$ be a dynamical system on X^α. A Lyapunov function is a continuous, real-valued function $V : X^\alpha \to \mathbb{R}$ such that*

$$\dot{V}(\phi) = \limsup_{t \to 0^+} \frac{V(S(t)\phi) - V(\phi)}{t} \leq 0$$

for all $\phi \in X^\alpha$.

The next theorem is a classical result from invariance theory and will be the main tool in the proof of point dissipativeness.

THEOREM 3.2. *Suppose that $u_0 \in X^\alpha$ and $\{S(t)u_0,\ t \geq 0\}$ lies in a compact set in X^α, then $\omega(u_0)$ is nonempty, compact, invariant, connected, and dist$(S(t)u_0, \omega(u_0)) \to 0$ as $t \to +\infty$.*

The following result has a classical and simple proof but we present it to make sure that our Lyapunov function (see later in this section) is suitable.

THEOREM 3.3. *Let V be a Lyapunov function on X^α and define $E = \{\phi \in X^\alpha : \dot{V}(\phi) = 0\}$, $\mathcal{M}$ the maximal invariant subset of E. If $\{S(t)u_0,\ t \geq 0\}$ lies in a compact set in X^α, then $S(t)u_0 \to \mathcal{M}$ as $t \to +\infty$.*

Proof. By hypothesis, $V(S(t)u_0)$ is nonincreasing for $t \geq 0$ and is bounded below (since orbits of points are precompact) so that $\ell = \lim_{t \to \infty} V(S(t)u_0)$ exists. If $y \in \omega(u_0)$, then $V(y) = \ell$, so also $V(S(t)y) = \ell$, $t \geq 0$, and so $\dot{V}(y) = 0$. Thus $\omega(u_0) \subset E$, so $\omega(u_0) \subset \mathcal{M}$ and the result is proved. □

We will apply these results and the results of §2 to obtain the existence of a global attractor for systems of reaction-diffusion equations without dispersion, that is, we consider the problem

$$(1.1)'' \qquad \begin{cases} u_t = D\Delta u - \gamma u + f(u), & \text{in } \Omega, \\ \dfrac{\partial u}{\partial n} = 0, & \text{in } \partial\Omega, \end{cases}$$

where $u = (u_1, u_2, \ldots, u_N)^\top$, $N \geq 1$, $D = \operatorname{diag}(d_1, \ldots, d_N)$, $d_i > 0$, $1 \leq i \leq N$, and $\gamma > 0$. The nonlinearity $f = (f_1, \ldots, f_N)^\top : \mathbb{R}^N \to \mathbb{R}^N$ is assumed to be a C^1 function that satisfies (1.3) and

$$(3.1) \qquad \frac{\partial f_i(u)}{\partial u_j} = \frac{\partial f_j(u)}{\partial u_i} \quad \forall u \in \mathbb{R}^N.$$

To prove that the solution operator for $(1.1)''$ is point dissipative, we must first prove that the orbit of a point $\phi \in X^\alpha$ is a compact subset of X^α. This is a consequence of the following result.

THEOREM 3.4. *In the problem (2.1) assume that the nonlinearity f^e is Lipschitz continuous in bounded subsets of X^α and that A is a sectorial operator with compact*

resolvent. If $T(t)\phi$ is a solution of (1.1)$''$ on $[0,\infty)$ with $\|T(t)\phi\|_\alpha$ bounded as $t \to \infty$, then $\{T(t)\phi,\ t \geq 0\}$ is a compact subset of X^α. Furthermore, if B is a bounded subset of X^α and $T(t)B$ remains in a bounded subset of X^α as $t \to \infty$, then $\{T(t)B,\ t \geq 1\}$ is a compact subset of X^α.

The proof of the above result is very simple and can be easily adapted from Henry [19, Thm. 3.3.6].

Now since all the hypotheses of Theorem 3.4 are satisfied for problem (1.1)$''$, we can conclude that orbits of points under the flow defined by (1.1)$''$ are compact subsets of X^α. From Theorem 3.2 the ω limit set of any point u_0 is a nonempty, compact, invariant set that attracts u_0 under the flow defined by (1.1)$''$.

From Theorem 3.3 we know that the set $\mathcal{M}$ attracts points of X^α. We need to find a Lyapunov function V for which E is a bounded subset of X^α; point dissipativeness will follow.

Let $V : X^\alpha \to \mathbb{R}$ be the function defined by

$$V(\phi) = \frac{1}{2}\int_\Omega \langle D\nabla\phi, \nabla\phi\rangle dx + \frac{\gamma}{2}\int_\Omega |\phi|^2 dx - \int_\Omega F(\phi)dx,$$

where $F : \mathbb{R}^N \to \mathbb{R}$ is such that $\nabla_u F(u) = f(u)$. Then V is continuous and

$$\dot{V}(\phi) \leq 0 \quad \forall \phi \in X^\alpha.$$

Therefore, we must prove that $E = \{\phi \in X^\alpha : \dot{V}(\phi) = 0\}$ is a bounded set in X^α.

The set E is the set of equilibrium points of (1.1)$''$ and therefore any function $\phi \in E$ *must satisfy* $\phi \in X^\alpha$ and

$$\text{(3.2)} \qquad \begin{cases} D\Delta\phi - \gamma\phi + f(\phi) = 0 & \text{in } \Omega, \\ \frac{\partial\phi}{\partial n} = 0 & \text{in } \partial\Omega. \end{cases}$$

To prove that E is a bounded subset of X^α we proceed in the following way. First we prove that there exists a constant $c > 0$ such that $\|\phi\|_{L^\infty(\Omega,\mathbb{R}^N)} \leq c\ \forall \phi \in E$ and then we use (3.2) to prove that $\|D\Delta\phi\|_{L^\infty(\Omega,\mathbb{R}^N)} \leq \max_{|s|\leq c} |-\gamma s + f(s)|$. The result will follow.

LEMMA 3.5. *There exists a constant $\xi > 0$ such that $\|\phi\|_{L^\infty(\Omega,\mathbb{R}^N)} \leq \xi$ for every $\phi \in E$.*

Proof. Let ξ_i be such that $sf(u_1,\ldots,u_{i-1},s,u_{i+1},\ldots,u_N) < 0\ \forall s$ such that $|s| \geq \xi_i$. Then suppose that $\phi := (\phi_1,\ldots,\phi_N) \in E$ and that $\max_{x\in\bar{\Omega}} \phi_k(x) = \phi_k(y) \geq \xi_k$ for some k. Thus, at y

$$\phi_k d_k \Delta\phi_k - \gamma\phi_k^2 + \phi_k f_k(\phi) = 0$$

and $\Delta\phi_k > 0$ since $\xi_k > 0$, but $\Delta\phi_k(y) \leq 0$ (if $y \notin \Omega$ we proceed as in Theorem 2.3). This is a contradiction and $\max_{x\in\bar{\Omega}} \phi_k(x) \leq \xi_k$, $1 \leq k \leq N$. In the same way we obtain that $\min_{x\in\bar{\Omega}} \phi_k(x) \geq -\xi_k$, $1 \leq k \leq N$, and the result is proved. □

COROLLARY 3.6. *The set E is a compact subset of X^α.*

Remark. It is very important, for some applications, to be able to obtain some a priori estimates that do not depend on the size of the diffusion coefficient. See, for example, Carvalho [3] and Carvalho and Oliveira [5]. This is an advantage that this technique (i.e., the use of invariant regions) has over the techniques employed in Hale [14, p. 77], even in the case $\alpha = \frac{1}{2}$. The estimates obtained there for the size of E strongly depend on the size of the diffusion coefficient. It is also important to obtain

good $L^\infty(\Omega)$ bounds on the attractor. Since models are an approximation of a real phenomenon, it is important to know that at least in a neighborhood of the attractor the approximation must be as accurate as we can get; outside this set it does not matter much.

COROLLARY 3.7 (Theorem 1.2). *The solution operator* $\{T(t),\ t \geq 0\}$ *for* (1.1)″ *is point dissipative and therefore has a global attractor.*

The proof of point dissipativeness follows from Theorem 3.3 and Corollary 3.6 and the proof of existence of a global attractor follows from the results in this section, §2, and Theorem 3.4.6 in Hale [14].

4. Proof of Theorem 1.3. In this section we consider the possibility that the problem (1.1) has the structure that the so-called gradient systems have. Such a special class of dynamical systems, for which the flow on the attractor can be better understood, are considered, for example, in Hale [14].

DEFINITION 4.1. *Let* Y *be a Banach space. A strongly continuous* C^r *semigroup* $T(t): Y \to Y$, $t \geq 0$, $r \geq 0$, *is said to be a gradient system if*

1. *Each bounded positive orbit is precompact.*

2. *There exists a Lyapunov function for* $T(t)$; *that is, there is a continuous function* $V: Y \to \mathbb{R}$ *with the property that*

(i) $V(y)$ *is bounded below,*

(ii) $V(u) \to \infty$ *as* $\|y\|_Y \to \infty$,

(iii) $V(T(t)y)$ *is nonincreasing in* t *for each* $y \in Y$,

(iv) *If* y *is such that* $T(t)y$ *is defined for* $t \in \mathbb{R}$ *and* $V(T(t)y) = V(y)$ *for* $t \in \mathbb{R}$, *then* y *is an equilibrium point.*

Observe that the Lyapunov function defined in §3 does not satisfy property 2 (ii). However, we still obtain that the dynamics in the attractor for (1.1) can be described as well as the dynamics of a gradient system.

To obtain these results we consider an auxiliary system, namely

$$\text{(4.1)} \qquad \begin{cases} u_t = D\Delta u - \gamma u + \tilde{f}(u) & \text{in} \quad \Omega, \\ \dfrac{\partial u}{\partial n} = 0 & \text{in} \quad \partial\Omega, \end{cases}$$

where $\tilde{f}$ is obtained from f in the following way. The attractor $\mathcal{A}$ for (1.1) in X^α is a bounded subset of $L^\infty(\Omega, \mathbb{R}^N)$ and therefore we can cut the nonlinearity f in such a way that all the properties of f are preserved. In addition, $\tilde{f}$ is globally Lipschitz continuous and the dynamics in the attractor remain unchanged. That is, the problem (4.1) has a global attractor in X^α which coincides with $\mathcal{A}$.

The problem (4.1) is well posed in $X^{1/2}$ and has an attractor $\tilde{\mathcal{A}}$ in $X^{1/2}$. Therefore we must have $\mathcal{A} \subset \tilde{\mathcal{A}}$. Since $\tilde{\mathcal{A}} \subset X^\alpha$ and we have that it is invariant, it must be contained in $\mathcal{A}$ and they are the same. This reasoning proves the following result.

THEOREM 4.2. *If* $\{T(t), t \geq 0\}$ *is the semigroup defined by* (1.1) *in* X^α, $\mathcal{A}$ *denotes its attractor, and* E *the set of equilibrium points, then* $\mathcal{A} = W^u(E) = \{y \in X^\alpha : T(-t)y$ *is defined for* $t \geq 0$ *and* $T(-t)y \to E$ *as* $t \to \infty\}$. *If, in addition, every element of* E *is hyperbolic, then* E *is a finite set and* $\mathcal{A} = \cup_{x \in E} W^u(x)$.

COROLLARY 4.3 (Theorem 1.3). *Let* $\Sigma_i = [-\bar{\xi}_i, \xi_i]$, *where* $\xi_i, \bar{\xi}_i$ *are positive constants,* $1 \leq i \leq N$. *If* $s_i f_i(u_1, \ldots, u_{i-1}, s, u_{i+1}, \ldots, u_N) < 0$ *whenever* $s_i \notin \Sigma_i$, *then* $\phi(x) \in \Sigma := \Sigma_1 \times \cdots \times \Sigma_N$, $\forall x \in \Omega$ *and* $\forall \phi \in \mathcal{A}$.

Proof. The proof of this result follows in two steps. First we observe that the equilibrium points satisfy that $\phi(x) \in \Sigma$, $\forall x \in \Omega$ (as in Lemma 3.5). The second

part follows from the fact that the α limit set of points in $\mathcal{A}$ are subsets of the set of equilibrium points and that any rectangle containing Σ is invariant. More specifically, let $\epsilon > 0$ and Σ_ϵ be an ϵ neighborhood of Σ. If $\phi \in \Sigma_\epsilon$, then there is a $\tau < 0$ and $e \in E$ such that $\|T(\tau)\phi - e\|_{L^\infty(\Omega,\mathbb{R}^N)} \leq K\|T(\tau)\phi - e\|_X^\alpha < \frac{\epsilon}{2}$, where K is the embedding constant of $X^\alpha \subset L^\infty(\Omega, \mathbb{R}^N)$. This implies that $(T(\tau)\phi)(x) \in \Sigma_\epsilon$, $\forall x \in \bar{\Omega}$. Since Σ_ϵ is invariant we have that $\phi \in \Sigma_\epsilon$.

This proves that for any $\epsilon > 0$ and $\phi \in \mathcal{A}$, $\phi(x) \in \Sigma_\epsilon$ $\forall x \in \bar{\Omega}$. The result follows. □

5. Systems arising from thin domains problems. In this section we consider a class of systems of weakly coupled parabolic partial differential equations of the form

$$\begin{cases} u_t = D\Delta u - \tilde{A}u + f(u) & \text{in } \Omega, \\ \dfrac{\partial u}{\partial n} = 0 & \text{in } \partial\Omega, \end{cases} \tag{5.1}$$

where $u = (u_1, u_2, \ldots, u_N)^\top$, $N \geq 1$, $D = \operatorname{diag}(d_1, \ldots, d_N)$, $d_i > 0$, $1 \leq i \leq N$ and the nonlinearity $f := (f_1(u), \ldots, f_N(u))^\top : \mathbb{R}^N \to \mathbb{R}^N$ is assumed to be a C^1 function that satisfies

$$u_i f_i(u) < 0 \qquad \forall u \in \mathbb{R}^N, \quad u_i \notin [\bar{\xi}, \xi], \quad 1 \leq i \leq N. \tag{5.2}$$

The matrix $\tilde{A}$ is taken as $\tilde{A} = M^{-1}B$ where $M = \operatorname{diag}(L_1, \ldots, L_N)$, with $0 \leq L_i \leq 1$ for $1 \leq i \leq N$ and B is the tridiagonal symmetric matrix

$$B = \begin{bmatrix} m_1 & r_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ r_1 & m_2 & r_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & r_2 & m_3 & r_3 & 0 & \cdots & 0 & 0 \\ 0 & 0 & r_3 & m_4 & r_4 & \cdots & 0 & 0 \\ \vdots & \vdots & & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & & & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & r_{N-3} & m_{N-2} & r_{N-2} & 0 \\ 0 & 0 & \cdots & 0 & 0 & r_{N-2} & m_{N-1} & r_{N-1} \\ 0 & 0 & \cdots & 0 & 0 & 0 & r_{N-1} & m_N \end{bmatrix}, \tag{5.3}$$

where

$$\begin{aligned} m_1 &= \tfrac{a_2}{2l_2} + \tfrac{a_1\rho}{\rho l_1 + a_1(1-\rho)}, \\ m_N &= \tfrac{a_N}{2l_N} + \tfrac{a_{N+1}\sigma}{\sigma l_{N+1} + a_{N+1}(1-\sigma)}, \\ m_k &= \tfrac{a_k}{2l_k} + \tfrac{a_{k+1}}{2l_{k+1}}, \quad k = 2, \ldots, N-1, \\ r_k &= \tfrac{-a_{k+1}}{2l_{k+1}}, \quad k = 1, \ldots, N-1, \end{aligned}$$

and $a_k > 0$, $l_k > 0$ for $1 \leq k \leq N+1$. This problem arises as a limiting problem for reaction-diffusion equations in thin domains around a point. (See Hale and Raugel [15], Carvalho [3], and Carvalho and Oliveira[5] for details.)

To prove that problem (5.1) has a global attractor we have to obtain an invariant region for equation (5.1) as in Theorem 2.3, find a Lyapunov function for it, and show that the set of equilibrium solutions are bounded in the $L^\infty(\Omega)$ norm as in §3.

Assume that f satisfies (3.1). Let $F : \mathbb{R}^N \to \mathbb{R}$ be such that

$$\nabla_u F(u) = f(u)$$

and $V : X^\alpha \to \mathbb{R}$ be the function defined by

$$V(\phi) = \int_\Omega \left(\sum_{i=1}^N \frac{d_i}{2} \|\nabla \phi_i\|^2 + \langle \tilde{A}\phi, \phi \rangle + F(\phi) \right) dx,$$

where $\langle \cdot, \cdot \rangle$ stands for the usual inner product in $\mathbb{R}^n$.

The function V is continuous and

$$\dot{V}(\phi) \leq 0 \quad \forall \phi \in X^\alpha.$$

Thus if $\phi \in E = \{\phi \in X^\alpha : \dot{V}(\phi) = 0\}$ we have that $\phi \in X^\alpha$ and

$$\begin{cases} D\Delta\phi - \tilde{A}\phi + f(\phi) = 0 & \text{in } \Omega, \\ \dfrac{\partial \phi}{\partial n} = 0 & \text{in } \partial\Omega. \end{cases} \tag{5.4}$$

We now prove the $L^\infty(\Omega)$ boundedness of the equilibrium solutions and the existence of the invariant regions for (5.1).

The following result is a consequence of the theory of invariant regions (see Smoller [24, p. 202]) and appears in Carvalho and Oliveira [5] when Ω is an interval.

LEMMA 5.1. *The rectangle* $[\bar{\rho}, \rho]^N$ *is an invariant region for* (5.1) *whenever* $\bar{\rho} \geq \bar{\xi}$ *and* $\rho \geq \xi$.

To prove that E is a bounded subset of X^α we proceed as in §3. First we prove that there exists a constant $c > 0$ such that $\|\phi\|_{L^\infty(\Omega, \mathbb{R}^N)} \leq c \quad \forall \phi \in E$ and then we obtain that

$$\|D\Delta\phi - \tilde{A}\phi\|_{L^\infty(\Omega, \mathbb{R}^N)} \leq \tilde{c} \quad \forall \phi \in E.$$

LEMMA 5.2. *If* E *is the set of equilibrium solutions of* (5.1) *there exists a constant* $\xi > 0$ *such that* $\|\phi\|_{L^\infty(\Omega)} \leq \xi$ *for every* $\phi \in E$.

Proof. Let ξ be such that $s_j f_j(s) < 0 \ \forall s$ such that $|s_j| \geq \xi$. Suppose that $\phi := (\phi_1, \phi_2, \ldots, \phi_N) \in E$ and let $\max_{x \in \bar{\Omega}} \phi_j(x) = \phi_j(y_j)$ (if $y_j \notin \Omega$ we proceed as in Theorem 2.3). We claim that $\phi_j(y_j) \leq \xi$ for all j.

Suppose $\phi_j(y_j) \leq \xi$ for $j = 1, 2, \ldots, k-1$ and $\phi_k(y_k) > \xi$ for some $1 \leq k \leq N$. From the kth equation we obtain

$$0 = \phi_k(y_k) d_k \Delta\phi_k(y_k) - \tfrac{r_k}{L_k} \phi_k(y_k)[\phi_{k+1}(y_k) - \phi_k(y_k)]$$

$$+ \tfrac{r_{k-1}}{L_k} \phi_k(y_k)[\phi_k(y_k) - \phi_{k-1}(y_k)] + \phi_k(y_k) f_k(\phi).$$

At y_k the maximum principle implies (see Protter and Weinberger [23, p. 65])

$$\phi_k(y_k) d_k \Delta\phi_k(y_k) \leq 0$$

and

$$\phi_k(y_k) f_k(\phi(y_k)) < 0.$$

Since $\phi_{k-1}(y_k) \leq \phi_{k-1}(y_{k-1}) \leq \xi$, we also have

$$\frac{r_{k-1}}{L_k}\phi_k(y_k)[\phi_k(y_k) - \phi_{k-1}(y_k)] < \frac{r_{k-1}}{L_k}\phi_k(y_k)[\xi - \phi_{k-1}(y_{k-1})] \leq 0.$$

These inequalities together with (5.5) imply that

$$-\frac{r_k}{L_k}\phi_k(y_k)[\phi_{k+1}(y_k) - \phi_k(y_k)] > 0.$$

It follows that $\phi_{k+1}(y_k) > \phi_k(y_k)$ and

$$\begin{cases} \phi_{k+1}(y_{k+1}) \geq \phi_{k+1}(y_k) > \phi_k(y_k) > \xi, \\ \phi_{k+1}(y_{k+1}) \geq \phi_{k+1}(y_k) > \phi_k(y_k) \geq \phi_k(y_{k+1}). \end{cases}$$

Using the same argument through the $(N-1)$th equation we obtain

$$\begin{aligned} &\phi_N(y_N) > \xi, \\ &\phi_N(y_N) > \phi_{N-1}(y_N), \end{aligned}$$

and the Nth equation gives

$$0 = \phi_N(y_N)d_N\Delta\phi_N(y_N) - \tfrac{m_N - r_{N-1}}{L_N}\phi_N(y_N)[0 - \phi_N(y_N)]$$

$$+\tfrac{r_{N-1}}{L_N}\phi_N(y_N)[\phi_N(y_N) - \phi_{N-1}(y_N)] + \phi_N(y_N)f_N(\phi(y_N)).$$

The same reasoning as before implies that

$$-\frac{m_N - r_{N-1}}{L_N}(\phi_N(y_N))^2 > 0.$$

This contradiction implies $\max_{x\in\bar{\Omega}} \phi_j(x) \leq \xi$ for $1 \leq j \leq N$. In the same way we obtain that $\min_{x\in\bar{\Omega}} \phi_j(x) \geq -\xi$ for $1 \leq j \leq N$ and the result is proved. □

COROLLARY 5.3. *Suppose that* (3.1) *and* (5.2) *hold. Then the set E is a bounded subset of X^α, the solution operator $\{T(t), t \geq 0\}$ for* (5.1) *is point dissipative and has a global attractor $\mathcal{A}$. In addition,*

$$\phi(x) \in [\bar{\xi}, \xi]^N \quad \forall x \in \bar{\Omega}$$

for all $\phi \in \mathcal{A}$ and

$$\mathcal{A} = W^u(E).$$

Furthermore, if each element of E is hyperbolic, E is finite and

$$\mathcal{A} = \cup_{x\in E} W^u(x).$$

If in addition

$$\frac{\partial f_i}{\partial u_j} > 0 \quad \text{for} \quad i \neq j, \tag{5.5}$$

then (5.1) is a cooperative system if Ω is a convex domain (see Kishimoto and Weinberger [20]). For such systems the following result holds.

PROPOSITION 5.4. *Let $\bar{u}$ be a nonconstant equilibrium solution of* (5.1). *Suppose that* (5.5) *holds on the range of $\bar{u}$. Then $\bar{u}$ is unstable.*

COROLLARY 5.5. *If* (5.5) *holds in Σ, then every nonconstant equilibrium solution for* (5.1) *is unstable; that is, if $\bar{u} \in E$ is stable then $\bar{u}$ is constant.*

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1971.

[2] H. BRÉZIS, *Analyse Fonctionnelle*, Masson Editeur, Paris, 1983.

[3] A. N. CARVALHO, *Infinite Dimensional Dynamics Described by Ordinary Differential Equations*, J. Diff. Equations, to appear; Notas do ICMSC, Série Matemática, 4 (1993).

[4] A. N. CARVALHO AND J. K. HALE, *Large diffusion with dispersion*, Nonlinear Anal., 17 (1991), pp. 1139–1151.

[5] A. N. CARVALHO AND L. A. F. OLIVEIRA, *Delay-Partial Differential Equations with Some Large Diffusion*, Nonlinear Anal., to appear; Notas do ICMSC, Série Matemática, 3 (1993).

[6] A. N. CARVALHO AND A. L. PEREIRA, *A Scalar Parabolic Equation Whose Asymptotic Behavior is Dictated by a System of Ordinary Differential Equations*, J. Diff. Equations, to appear.

[7] K. CHUEH, C. CONLEY AND J. SMOLLER, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.

[8] E. CONWAY, D. HOFF, AND J. SMOLLER, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1–16.

[9] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, vol.* 1, John Wiley and Sons, New York, 1989.

[10] A. FRIEDMAN, *Partial Differential Equations*, Krieger Publishing Company, Melbourne, FL, 1983.

[11] G. FUSCO, *On the explicit construction of an ODE which has the same dynamics as a scalar Parabolic PDE*, J. Differential Equations, 69 (1987), pp. 85–110.

[12] J. K. HALE, *Asymptotic Behavior and Dynamics in Infinite Dimensions*, Res. Notes Math. 132, Pitman Books Limited, London, pp. 1–42, 1985.

[13] ———, *Large diffusivity and asymptotic behavior in parabolic systems*, J. Math. Anal. Appl., 118 (1986), pp. 455–466.

[14] ———, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monographs 25, American Mathematical Society, Providence, RI, 1988.

[15] J. K. HALE AND G. RAUGEL, *Attractors for Dissipative Evolutionary Equations*, CDSNS - Report 72, Georgia Tech., Atlanta, GA 1992.

[16] J. K. HALE AND C. ROCHA, *Varying boundary conditions with large diffusivity*, J. Mat. Pures et Appl., 66 (1987), pp. 139–158.

[17] ———, *Interaction of diffusion and boundary conditions*, Nonlinear Anal., 11 (1987), pp. 633–649.

[18] J. K. HALE AND K. SAKAMOTO, *Shadow systems and attractors in reaction-diffusion equations*, Appl. Anal., 32 (1989), pp. 287–303.

[19] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lectures Notes in Math., 840, Springer-Verlag, Berlin, 1981.

[20] K. KISHIMOTO AND H. F. WEINBERGER, *The spatial homogeneity of stable equilibria of some reaction-diffusion systems in convex domains*, J. Differential Equations, 58 (1985), pp. 15–21.

[21] J. MOSER, *A sharp form of an inequality by N. Trudinger*, Indiana Univ. Math. J., 20 (1971), pp. 1077–1092.

[22] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci., 44, Berlin, Springer-Verlag, 1983.

[23] M. F. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, New York, 1967.

[24] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Grundlehren Math. Wiss., 258, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 428–435, March 1995

009

ON STABILITY OF A DYNAMICAL SYSTEM*

CHARLES S.C. LIN†, BIN YANG†, AND FUDONG CHEN†

Abstract. This note solves an open problem raised by Zeeman in [*Nonlinearity*, 1 (1988), pp. 115–155]. It extends his results about the stability of a dynamical system from C^∞-vector fields to C^m-vector fields, where $1 \leq m \leq \infty$. For any C^m-vector field v, the existence, uniqueness, and global attraction of the steady state of the Fokker–Planck equation in C^{m-1} space is proved. The steady states are used as a tool to classify vector fields. The density of stable C^m-vector fields is also proved.

Key words. dynamical system, stability, semigroup, positive operators, the Fokker–Planck equation, compact manifold

AMS subject classifications. 34C35, 47D05, 58F10

1. Introduction.

Given a vector field v on an oriented n-dimensional Riemannian manifold and $\epsilon > 0$, the time-dependent Fokker–Planck equation is given as follows:

$$\frac{du}{dt} = \epsilon \Delta u - \nabla \cdot (uv), \tag{1}$$

$$u(0) = u_0. \tag{2}$$

Zeeman proved that there is a smooth steady state u of the Fokker–Planck equation for a smooth vector field v on a connected compact Riemannian manifold without boundary cf. [1]. A new definition of stability of a dynamical system is introduced via the steady state. The new definition has a number of advantages over structural stability. Zeeman gave several good examples to compare the two stabilities. One of the advantages of the new definition is that stable vector fields are dense in the C^∞ topology; cf. [1] and [2].

In this paper we deal with C^m-vector fields on an oriented compact Riemannian manifold without boundary, where $1 \leq m < \infty$. The main results are as follows:

(i) For any C^m-vector field v, there is a unique steady state u of the Fokker–Planck equation in C^{m-1} space. Furthermore, all solutions tend to the steady state u.

(ii) The stable C^m-vector fields are dense in the space of C^m-vector field.

2. The steady state of the Fokker–Planck equation.

Let X be an oriented n-dimensional Riemannian manifold, and X connected compact without boundary. In what follows, we shall fix a chart on X. Let $k = (k_1, k_2, \ldots, k_n)$, $|k| = k_1 + k_2 + \cdots + k_n$, where k_i are integers for $i = 1, 2, \ldots, n$. Let C^m be the set of C^m-differentiable functions on X and $C^m(X)$ the space of C^m function with C^m topology defined on a chart by

$$\| u \|_m = \max \sum_{0 \leq |k| \leq m} \left| \frac{\partial^{|k|}}{\partial x^k} u \right|.$$

* Received by the editors October 8, 1992; accepted for publication (in revised form) September 14, 1993.

† Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, Illinois 60680.

Let

$$U^m = \left\{ u \in C^m; u \geq 0 \text{ and } \int_X u = 1 \right\}$$

and V^m be the space of all C^m-vector fields on X with C^m topology.

DEFINITION 1. *A solution $u^v(t)$ of equation (1) is called a* steady state *if $\frac{d}{dt}u^v(t) = 0$ and $u^v(t) \in U^{m+1}$.*

Given a C^m-vector field v and $\epsilon > 0$, we define the Fokker–Planck linear operator FP_v in $C^{m-1}(X)$ as follows:

$$FP_v u = \epsilon \Delta u - \nabla \cdot (uv), \tag{3}$$

$$D(FP_v) = \{u \in C^{m-1}(X); FP_v u \in C^{m-1}(X)\}. \tag{4}$$

We shall first recall some definitions needed for our results. A *cone* K in a Banach space Y is a closed subset of Y such that for any $\alpha > 0, \alpha K \subseteq K, K \cap (-K) = \{0\}$ and $K + K \subseteq K$.

A *positive operator* P (with respect to cone K) in Y is an operator in Y such that $P(K) \subseteq K$.

A positive operator P is called *strongly positive* if P maps every nonzero point in K onto the interior of K.

The following facts are known; cf. [1] and [3].

(i) There is a semigroup $G(t)$ of bounded linear operator in $C^{m-1}(X)$ with generator FP_v, i.e., the Fokker–Planck equation (1) has a solution $G(t)u_0$ for any $u_0 \in D(FP_v)$.

(ii) $G(t)$ has an expression of the form

$$G(t)u(x) = \int_X g(x, y, t)u(y)\, dy, \tag{5}$$

$$\int_X g(x, y, t)\, dx = 1, \tag{6}$$

where $g(x, y, t)$ is a solution of the Fokker–Planck equation (1) with initial condition $\delta(x - y)$.

(iii) $G(t)$ is a strongly positive operator on $C^{m-1}(X)$ and $G(t)U^{m+1} \subseteq U^{m+1}$ for any $t > 0$.

Therefore, in order to obtain the existence, uniqueness, and global attraction of the steady state of the Fokker–Planck equation, it is enough to prove the following theorem (cf. [1], [2], and [4]).

THEOREM 2.1. *The operator $G(t)$ is compact on $C^{m-1}(X)$ for all $t > 0$.*

To prove Theorem 2.1, we need several lemmas.

LEMMA 2.2. *Let $B = \{f \in C^m(X), \| f \|_m = 1\}$. An operator G is compact on $C^m(X)$ if $\{\frac{\partial^{|k|}}{\partial x^k}(Gf); f \in B\}$ is equicontinuous for all $0 \leq |k| \leq m$ and $\{\partial^{|k|}/\partial x^k(Gf)(x); f \in B\}$ is relatively compact in R for all x in a chart and $0 \leq |k| \leq m$.*

Proof. Applying Ascoli's theorem for each index k, we have a convergent subsequence $(\partial^{|k|}/\partial x^k)(Gf_{k_n}), n = 1, 2, \ldots$. Without loss of generality, by selecting a subsequence of a subsequence if necessary, we may assume there is a subsequence $(\partial^{|k|}/\partial x^k)(Gf_n)$ which is convergent in $C(X)$ for any index k. That implies Gf_n is a Cauchy sequence in $C^m(X)$. Hence, G is a compact operator on $C^m(X)$.

Let $h : R \to R$ be given by $h(r) = e^{-r^2/2}$. It is easy to check that $h^{(l)}(r) = h(r)p_l(r)$, where $p_l(r)$ is a polynomial of degree l. Therefore, Lemma 2.3 is trivial.

LEMMA 2.3. *There is a constant $M > 0$ such that*

$$\max_{r \in R} |h^{(l)}(r)| \leq M$$

and locally

$$|h^{(l)}(r) - h^{(l)}(r')| \leq M|r - r'|$$

for all $0 \leq l \leq m$.

LEMMA 2.4. *For any $v \in V^m$ and $\epsilon > 0$, we have*

$$\left| \frac{\partial^{|k|} g}{\partial x^k}(x, y, t) - \frac{\partial^{|k|} g}{\partial x^k}(x', y, t) \right| \leq M_1(t)|x - x'|, \qquad 0 \leq |k| \leq m \tag{7}$$

for $t > 0$, where $g(x, y, t)$ is a solution of the Fokker–Planck equation arising from v with initial condition $\delta(x - y)$, and $M_1(t)$ is a continuous function.

Proof. Let $X = R^n, v_0 = c - \kappa x$, where c, κ are constants, $c \in R^n, \kappa \in R$. Let $g_0(x, y, t)$ be the solution of the Fokker–Planck equation for v_0 with initial condition the Dirac function $\delta(x - y)$. Then (cf. [1] and [3])

$$g_0(x, y, t) = (2\pi\sigma)^{-n/2} e^{-|x-y-\mu|^2/2\sigma},$$

where $x, y \in R^n$,

$$\sigma = \begin{cases} \frac{\epsilon}{\kappa}(1 - e^{-2\kappa t}) & \text{if } \kappa \neq 0, \\ 2\epsilon t & \text{if } \kappa = 0, \end{cases}$$

and

$$\mu = \begin{cases} c(1 - e^{-\kappa t}) & \text{if } \kappa \neq 0, \\ ct & \text{if } \kappa = 0. \end{cases}$$

By Lemma 2.3, we have

$$\left| \frac{\partial^{|k|} g_0}{\partial x^k}(x, y, t) - \frac{\partial^{|k|} g_0}{\partial x^k}(x', y, t) \right| \leq M(2\pi)^{-n/2} \sigma^{-(n+1)/2} |x - x'|, \qquad 0 \leq |k| \leq m. \tag{8}$$

Since any manifold X can be locally approximated by a Euclidean space and any vector field v on R^n has linear approximation at the origin, we have by arguments similar to [1] that

$$\left| \frac{\partial^{|k|} g}{\partial x^k}(x, y, t) - \frac{\partial^{|k|} g}{\partial x^k}(x', y, t) \right| \leq M_1|x - x'|, \qquad 0 \leq |k| \leq m \tag{9}$$

for the solution $g(x, y, t)$ of the Fokker–Planck equation arising from v with initial condition $\delta(x - y)$, where M_1 depends on t. Obviously, the compactness of X is necessary here.

Proof of Theorem 2.1. Since

$$G(t)f(x) = \int_X g(x, y, t) f(y)\, dy, \tag{10}$$

and $(\partial^{|k|}/\partial x^k)g(x,y,t)$ is continuous for $0 \le |k| \le m-1$, we have

$$\frac{\partial^{|k|}}{\partial x^k} G(t)f(x) = \int_X \frac{\partial^{|k|}}{\partial x^k} g(x,y,t)f(y)\,dy.$$

Thus, by Lemma 2.4 and Lemma 2.2, we see that $G(t)$ is a compact operator on $C^{m-1}(X)$.

COROLLARY 2.5. *$G(t)$ has a unique fixed point u in $C^{m-1}(X)$ with $u \ge 0$ and $\int_X u = 1$. Furthermore, $C^{m-1}(X)$ has a decomposition*

$$C^{m-1}(X) = E + H, \quad E \cap H = \{0\},$$

where

$$E = \operatorname{span}\{u\},$$
$$H = \{h \in C^{m-1}(X);\ G^n(t)h \to 0 \text{ as } n \to \infty\},$$

or

$$H = \left\{h \in C^{m-1}(X);\ \int_X h = 0\right\}.$$

Proof. Since $G(t)$ is a compact and strongly positive operator on $C^{m-1}(X)$, spectral radius r of $G(t)$ is a simple eigenvalue associated with an eigenfunction $u \ge 0$. Furthermore, $C^{m-1}(X)$ can be written as the sum of $G(t)$-invariant subspaces

$$C^{m-1}(X) = E + H, \qquad E = \operatorname{span}\{u\},$$

and $G(t)|_H$ has spectral radius $< r$.

We may assume $\int u = 1$ by scaling u. Hence, $r = 1$ follows from the fact that $\int G(t)u = \int u = 1$, u is a fixed point of $G(t)$ in $C^{m-1}(X)$. Suppose $G(t)$ has another fixed point u_1 in $C^{m-1}(X)$ with $u_1 \ge 0$ and $\int u_1 = 1$, then $u_1 = cu$. $\int u_1 = \int u = 1$ implies $c = 1$.

It is easily seen that

$$H \subset \{h \in C^{m-1}(X);\ G^n(t)h \to 0 \text{ as } n \to \infty\},$$

since the spectral radius of $G(t)|_H$ is less than 1. Conversely suppose $f \in C^{m-1}(X)$ and $G^n(t)f \to 0$ as $n \to \infty$. Since $f = e + h, e \in E, h \in H$ and $G(t)e = e, G^n(t)h \to 0$ as $n \to \infty$, we have $e = 0$ and $f = h \in H$. Consequently,

$$H = \{h \in C^{m-1}(X);\ G^n(t)h \to 0 \text{ as } n \to \infty\}.$$

By (5) and (6), $\int G^n(t)f = \int f$ for any $f \in C^{m-1}(X)$. If $h \in H$ then $G^n(t)h \to 0$ as $n \to \infty$,by the arguments above. Therefore $\int G^n(t)h \to 0$ as $n \to \infty$. It follows that $\int h = 0$.

On the other hand, suppose $f \in C^{m-1}(X)$ and $\int f = 0$. Let $f = e + h, e \in E$ and $h \in H$, then $e = cu$. Since $\int h = 0$ as shown above, $c = \int cu = \int e = \int(f - h) = 0$. Hence, $e = 0$ and $f = h \in H$. This completes the proof of Corollary 2.5.

THEOREM 2.6. *For any vector field $v \in V^m$, the Fokker–Planck equation for v with ϵ-diffusion has a unique steady state $u \in U^{m+1}$, and all solutions tend to the steady state u.*

Proof. Let $\tau > 0$ and u be the unique fixed point of $G(\tau)$ in Corollary 2.5. Since

$$G(\tau)G(t)u = G(t)G(\tau)u = G(t)u,$$

$$G(t)u \geq 0 \text{ and } \int G(t)u = \int u = 1$$

for any $t > 0$, we have $G(t)u = u$ for any $t > 0$ by Corollary 2.5. Hence

$$\frac{d}{dt}(G(t)u) = \frac{d}{dt}(u) = 0,$$

$u \in D(FP_v)$ and u is a steady state of the Fokker–Planck equation in U^{m+1}.

If there is another steady state u_1, then $u_1 \in U^{m+1}$ and

$$\frac{d}{dt}(u_1) = FP_v u_1 = 0.$$

That implies u_1 is an eigenvector of FP_v associated with eigenvalue 0. Hence, u_1 is an eigenvector of $G(t)$ for any $t > 0$ corresponding to the eigenvalue 1 (cf. [5]). In particular,

$$G(\tau)u_1 = u_1.$$

The uniqueness of Corollary 2.5 implies $u = u_1$.

For any solution $G(t)u_0$ in U^{m+1}, $G(t)u_0 - u \in H$ follows from $\int (G(t)u_0 - u) = 0$ and Corollary 2.5. Since $\{G(t)u_0;\ 0 \leq t \leq \tau\}$ is bounded, $\{G(t)u_0 - u;\ 0 \leq t \leq \tau\}$ is bounded also, say, by M.

By the strong positivity of $G(\tau)$, the spectral radius r of $G(\tau)|_H$ is less than one. We can select a ρ such that $r < \rho < 1$. Then, there is an n_0 such that $\| G^n(\tau)|_H) \| \leq \rho^n$ for all $n > n_0$. Given $\epsilon > 0$, there is an n_1 such that $n_1 > n_0$ and $\rho^n M < \epsilon$ for all $n > n_1$. Hence, for any $t > n_1\tau$, $t = n\tau + s$, where $n > n_1$ and $0 < s < \tau$, we have

$$\begin{aligned} \| G(t)u_0 - u \|_{m-1} &= \| G^n(\tau)G(s)u_0 - G^n(\tau)u \|_{m-1} \\ &\leq \rho^n \| G(s)u_0 - u \|_{m-1} \\ &\leq \rho^n M \leq \epsilon. \end{aligned}$$

This completes the proof of Theorem 2.6

3. Density of stable vector fields. In this section we prove that stable vector fields are dense in the space of all C^m-vector fields on a compact connected Riemannian manifold.

DEFINITION 2. *Two C^∞-functions $u, u' : X \to R$ are said to be* equivalent *if there exist diffeomorphisms α, β of X, R, respectively, such that the following diagram commutes.*

$$\begin{array}{ccc} X & \xrightarrow{u} & R \\ \downarrow \alpha & & \downarrow \beta \\ X & \xrightarrow{u'} & R \end{array}$$

Two C^m-functions $u, u' : X \to R$ are said to be C^m-equivalent *if either $u = u'$ or there are neighborhoods of u and u', say, O_u and $O_{u'}$ respectively, in $C^m(X)$ such that every function in $O_u \cap C^\infty$ is equivalent to every function in $O_{u'} \cap C^\infty$.*

A C^m-function f is called C^m-stable *if f has a neighborhood of C^m-equivalents in $C^m(X)$.*

DEFINITION 3. *Two C^m-vector fields v, v' on X are said to be* ϵ-equivalent *in V^m if their corresponding steady states of the Fokker–Planck equation with ϵ-diffusion are C^{m-1}-equivalent in $C^{m-1}(X)$.*

A C^m-vector field v is called ϵ-stable *if it has a neighborhood O in V^m such that every pair of vector fields in O are ϵ-equivalents.*

A C^m-vector field is called stable *if it is ϵ-stable for arbitrarily small $\epsilon > 0$.*

Define a map

$$\pi^\epsilon : V^m \to U^{m+1}$$

by assigning to each $v \in V^m$ the steady state $\pi^\epsilon(v) = u^{v,\epsilon}$ of the Fokker–Planck equation for v with ϵ diffusion. By Corollary 2.5 of §2, the map π^ϵ is well defined. We will use different topologies on the set U^{m+1} in Theorems 3.1 and 3.2.

THEOREM 3.1. *The map $\pi^\epsilon : V^m \to (U^{m+1}, \|\cdot\|_{m-1})$ is differentiable and hence continuous.*

Proof. Recall that

$$C^{m-1}(X) = E + H, \qquad E \cap H = \{0\},$$

where E is the one-dimensional subspace spanned by the steady state, and H is the complementary invariant subspace of $G(t)$. Since $G(t)$ is the semigroup generated by FP_v, H is also an invariant subspace under FP_v. Define

$$\Gamma = FP_v|_H.$$

Γ is an infinitesimal generator of $G(t)|_H$. By the spectral mapping theorem (cf. [5]), 0 is a resolvent point of Γ since 1 is a resolvent point of $G(t)|_H$. Therefore Γ^{-1} is continuous.

Given $v \in V^m$, let $u = \pi^\epsilon(v)$. For any $v' \in V^m, \nabla \cdot (uv') \in C^{m-1}(X)$. Furthermore, $\nabla \cdot (uv') \in H$ because $\int \nabla \cdot (uv') = 0$ follows from integration by parts. Let $u' = \Gamma^{-1}(\nabla \cdot (uv'))$. Then

$$\nabla \cdot (uv') = \Gamma(u') = \epsilon \Delta u' - \nabla \cdot (u'v),$$

$$\begin{aligned} \epsilon \Delta(u + \delta u') &= \epsilon \Delta u + \delta(\epsilon \Delta u') \\ &= \nabla \cdot (uv) + \delta[\nabla \cdot (uv') + \nabla \cdot (u'v)] \\ &= \nabla \cdot [(u + \delta u')(v + \delta v')] + O(\delta^2) \end{aligned}$$

for any $\delta > 0$. It implies

$$\pi^\epsilon(v + \delta v') = u + \delta u' + O(\delta^2).$$

The map $v' \to u' = \Gamma^{-1}(\nabla \cdot (uv'))$ gives the derivative $T_v\pi^\epsilon$. It is continuous because it is the composition of three continuous maps:

$$v' \longmapsto uv' \longmapsto \nabla \cdot (uv') \longmapsto \Gamma^{-1}(\nabla \cdot (uv')),$$

the first being continuous because u is in C^{m+1}, the second being continuous because V^m has C^m topology but H has C^{m-1} topology and the third being continuous by

the continuity of Γ^{-1}. Therefore π^ϵ is differentiable and continuous from V^m to $(U^{m+1}, \|\cdot\|_{m-1})$.

THEOREM 3.2. *The map π^ϵ is open from V^m to $(U^{m+1}, \|\cdot\|_{m+1})$.*

Proof. Let

$$G = \{g \in V^m;\ g = -\nabla f,\ \text{for some } f \in C^{m+1}\},$$

$$W = \{w \in V^m;\ \nabla \cdot w = 0\}.$$

We define the following operators:

$$\pi_1 : G \times W \longrightarrow G, \qquad \pi_1(g, w) = g;$$
$$\tilde{\pi}\colon G \times W \longrightarrow V^m, \qquad \tilde{\pi}\,(g, w) = g + \frac{w}{u}, \quad \text{where } u = \pi^\epsilon(g);$$
$$\tilde{\psi}\colon V^m \longrightarrow G \times W, \qquad \tilde{\psi}\,(v) = (\epsilon \nabla u/u, uv - \epsilon \nabla u), \quad \text{where } u = \pi^\epsilon(v).$$
$$\psi : U^{m+1} \longrightarrow G, \qquad \psi(u) = -\nabla(-\epsilon \ln u);$$
$$\pi^\epsilon|_G : G \longrightarrow U^{m+1}, \qquad \pi^\epsilon|_G(g) = \pi^\epsilon(g) = e^{-f/\epsilon} \quad \text{for some } -\nabla f = g.$$

Then it is easily checked by computation that $\tilde{\pi}$ and $\tilde{\psi}$ are inverse homeomorphisms between $G \times W$ and V^m and $\pi|_G$ and ψ are inverse homeomorphisms between G and $(U^{m+1}, \|\cdot\|_{m+1})$(cf. [1]). Furthermore, we have

$$\pi^\epsilon = (\pi^\epsilon|_G)\pi_1\,\tilde{\psi}\,.$$

Hence π^ϵ is open from V^m to $(U^{m+1}, \|\cdot\|_{m+1})$, because it is the composition of three open maps.

THEOREM 3.3. *For any $\epsilon > 0$, ϵ-stable C^m-vector fields are dense in V^m.*

Proof. Let U_0 be the set of C^{m-1}- stable functions in $(U^{m+1}, \|\cdot\|_{m-1})$. Let V_0 be the inverse image of U_0 under π^ϵ. Then V_0 is the set of ϵ-stable C^m-vector fields because π^ϵ is continuous. In fact, let $v \in V_0$, then $u = \pi^\epsilon v \in U_0$. Since u is C^{m-1}-stable, there is an C^{m-1}-equivalent neighborhood O_u of u in $C^{m-1}(X)$. Therefore, inverse image of O_u under π^ϵ is an ϵ-equivalent neighborhood of v.

Let U_{00} be the set of C^{m+1}-Morse functions with distinct critical values in $(U^{m+1}, \|\cdot\|_{m+1})$. Let V_{00} be the inverse image of U_{00} under π^ϵ. Then U_{00} is dense in U^{m+1} with C^{m+1} topology; cf. [4]. Furthermore, we have $U_{00} \subset U_0$ and $V_{00} \subset V_0$; cf. [6]. For any $v \in V^m$ and its neighborhood O_v, let $u = \pi^\epsilon v$ and $O_u = \pi^\epsilon O_v$. Then O_u is a neighborhood of u in $C^{m+1}(x)$ by Theorem 3.2. Therefore, there is a u_0 in $U_{00} \cap O_u$. We can select v_0 from $V_{00} \cap O_v$ such that $\pi^\epsilon v_0 = u_0$. This implies that ϵ-stable C^m-vector fields are dense in V^m.

THEOREM 3.4. *Stable C^m-vector fields are residual and therefore dense in V^m.*

Proof. It should be noted that ϵ-stable vector fields form an open subset in V^m for any $\epsilon > 0$. Let

$$W = \cap_{n=1}^{\infty} V_{1/n},$$

where $V_{1/n}$ is $(1/n)$-stable set. Then any vector field in W is stable. Now W is residual in V^m because it is a countable intersection of open dense subsets in a Baire space. Therefore the density of stable C^m-vector fields in V^m follows from the density of W in V^m.

REFERENCES

[1] E. C. ZEEMAN, *Stability of dynamical systems*, Nonlinearity, 1 (1988), pp. 115–155.
[2] ———, *Presidential address on the classification of dynamical systems*, Bull. London Math. Soc., 20 (1988), pp. 545–557.
[3] H. RISKEN, *Fokker–Planck Equation*, Springer-Verlag, Berlin, New York, 1984.
[4] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and Their Singularities*, Springer-Verlag, Berlin, New York, 1973.
[5] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Amer. Math. Soc. Colloq. Publ., Vol. 31, American Mathematical Society, Providence, RI, 1957.
[6] C. S. C. LIN, F. CHEN, AND B. YANG, *On generalized linear dynamical system and stability on* $Diff^m(X)$, submitted.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 436–445, March 1995

010

MULTIEXISTENCE OF SLOWLY OSCILLATING PERIODIC SOLUTIONS FOR DIFFERENTIAL DELAY EQUATIONS *

YULIN CAO†

Abstract. This paper presents a checkable condition on the function f such that the differential delay equation $\dot{x}(t) = -f(x(t-1))$ has at least n distinct slowly oscillating periodic solutions, where n is any natural number or infinity. As an example, an equation is demonstrated to satisfy the condition proposed for $n = +\infty$, and therefore, it has infinitely many slowly oscillating periodic solutions.

Key words. slowly oscillating periodic solutions, differential delay equations

AMS subject classification. 34K15

1. Introduction.

This paper discusses the differential delay equation

$$\dot{x}(t) = -f(x(t-1)), \tag{1.1}$$

where f is a continuously differentiable function and $f(0) = 0$. For any given number n (n is allowed to be infinity), we will give a checkable condition on the function f such that (1.1) has at least n distinct slowly oscillating periodic solutions. Here, a slowly oscillating periodic solution means that the distances between its zeros are greater than one (i.e., the delay time). Two slowly oscillating periodic solutions are distinct if they are not equal under any time-shift. In §4, an equation is constructed to satisfy the condition proposed for $n = +\infty$, and therefore, it has infinitely many slowly oscillating periodic solutions.

Equation (1.1) is primarily important in the study of slowly oscillating periodic solutions, not only because it is the simplest one among the differential delay equations (or the functional differential equations), but also because some other important differential delay equations can be changed into this form (1.1). For example, the nonlinear equation $\dot{y}(t) = -y(t-1)N(y(t))$ can be changed into (1.1). (See [13, pp. 276–277].) This nonlinear equation has several applications, and the question of multiexistence of slowly oscillating periodic solutions is pertinent here. Cunningham [5] has used it as a population model, Wright [20] has used it in the theory of asymptotic prime number density, and Jones [9] has used it to describe a control system.

Nussbaum [17, Thm. 2.2] has proved that there is only one slowly oscillating periodic solution of the equation

$$\dot{x}(t) = -\alpha h(x(t-1)), \qquad \alpha > \frac{\pi}{2}, \tag{1.2}$$

under the assumptions (i) h is an odd and continuously differentiable function, (ii) $h'(0) = 1$ and $h'(x) > 0$ is monotonically decreasing for $x > 0$, and (iii) $\phi(x) = h(x)x^{-1}$ is strictly monotonically decreasing for $x > 0$. He has demonstrated a special equation which violates the assumption (iii) (he has assumed that $\phi(x)$ is only monotonically decreasing for $x > 0$) and has more than one slowly oscillating periodic solution. Our example in Theorem 4.1 shows that, if the assumption (iii) is violated, there could be

* Received by the editors September 19, 1990; accepted for publication (in revised form) October 11, 1993.

† Department of Mathematics, University of Georgia, Athens, Georgia 30602.

infinitely many slowly oscillating periodic solutions of (1.2). Our main result is based on Kaplan and Yorke's Trajectory Crossing Lemma. (See [11] and [12].) Our basic idea is to construct some solutions whose orbits in the $(x(t), x(t-1))$ plane spiral inward or outward and do not cross each other. If a pair of those orbits forms an annulus-like region, then there is a slowly oscillating periodic solution whose orbit is in this region. Consequently, if we can construct n such disjoint annulus-like regions, then there are n distinct slowly oscillating periodic solutions. This idea is similar to the Poincare–Bendixson Annular Region Theorem for the existence of a limit cycle for planary ordinary differential equations.

2. Definitions and primary results. First, we introduce some definitions and results from [12]. Consider the scalar differential delay equation

$$\dot{x}(t) = -f(x(t-1)), \tag{2.1}$$

where the function $f(x)$ is continuously differentiable and satisfies the following assumption:

$$f(0) = 0, \qquad f'(x) > 0 \quad \text{for all } x \in \mathbb{R}. \tag{H1}$$

Let $C = C([-1,0], \mathbb{R})$ be the Banach space of all continuous functions from $[-1,0]$ to $\mathbb{R}$. Define C_* by the set of all φ in C satisfying

(i) φ has at most one zero in $[-1,0]$, and

(ii) φ must change sign at any zero in $(-1,0)$.

Write $x(t) = x(t; t_0, \varphi)$ to denote the solution of (2.1) satisfying $x(t_0 + \theta) = \varphi(\theta)$ for $\theta \in [-1,0]$. Denote $x_t \in C$, as usual, by $x_t(\theta) = x(t+\theta)$ for $\theta \in [-1,0]$. It is easy to see that $x(t)$ exists for all $t \geq t_0$.

LEMMA 2.1. *Suppose* (H1) *is satisfied. If* $\varphi \in C_*$, *then* $x_t \in C_*$ *for all* $t \geq t_0$, *and the zeros of* x *and* $\dot{x}$ *alternate on* $(t_0, +\infty)$. *If* x *has only finitely many zeros in* $(t_0, +\infty)$, *then* $x(t)$ *monotonically goes to zero eventually, as* $t \to +\infty$.

Proof. The first part is given by [11, Prop. 2.1]. The second part can be deduced from (2.1), because $x(t-1) > 0 (x(t-1) < 0)$ implies that $\dot{x}(t) < 0 (\dot{x}(t) > 0)$.

From Lemma 2.1, one can see that, if $\varphi \in C_*$, then $x(\cdot; t_0, \varphi)$ is a slowly oscillating solution of (2.1), that is, the distance between any pair of successive zeros of x is greater than the delay time (i.e., one). Also from Lemma 2.1, we know that $x(t)$ and $x(t-1)$ cannot equal zero at the same t. The basic idea of [11] is to compare slowly oscillating solutions in the $(x(t), -x(t-1))$ plane. □

DEFINITION 2.1. *Given* $\gamma = (\gamma_1, \gamma_2)$ *and* $S = (s_1, s_2)$ *in* $\mathbb{R}^2$, *we say that* γ *is outside* S *if* $\gamma \neq S$ *and if either of the following is satisfied*:

(i) $\gamma_1 = s_1, |\gamma_2| > |s_2|$, *and* $\gamma_2 s_2 > 0$,

(ii) $s_2 = 0, |\gamma_1| \geq |s_1|$, *and* $\gamma_1 s_1 > 0$.

Let $\gamma(t) = (\gamma_1(t), \gamma_2(t))$ *and* $S(t) = (s_1(t), s_2(t))$ *be two parametrized continuous curves in* $\mathbb{R}^2$ *for* t *in intervals* I_γ *and* I_s, *respectively. We say that* $\gamma | I_\gamma$ *is outside* $S | I_s$ *if there exists a nondecreasing continuous function* T *from* I_γ *onto* I_s *such that* $\gamma(t)$ *is outside* $S(T(t))$ *for all* $t \in I_\gamma$.

Observe that, when the point S in the (u,v) plane is above the u axis, "outside S" means straight above S; when S is below the u axis, "outside S" means straight below S; when S is on the right u axis, "outside S" means on the right side of S; when S is on the left u axis, "outside S" means on the left side of S. Assuming x and y are two solutions of (2.1), whenever we say that $x|I_x$ is outside $y|I_y$, we mean that the curve γ defined by $\gamma(t) = (x(t), -x(t-1))$, for t in the interval I_x, is outside the curve S

defined by $S(t) = (y(t), -y(t-1))$ for t in the interval I_y. When $\sup I_x = +\infty$, we call γ an orbit of x in $\mathbb{R}^2$. From Lemma 2.1 and the discussion that followed, one can see that the orbit of a slowly oscillating solution x of (2.1) rotates in the $(x(t), -x(t-1))$ plane clockwise around the origin. The next lemma is a modification of Kaplan and Yorke's Trajectory Crossing Lemma in [11] and [12].

LEMMA 2.2 (Trajectory Crossing Lemma). *Suppose* (H1) *is satisfied. Let x and y be two solutions of* (2.1). *Assume that $x_{t_0}, y_{T_0} \in C_*$, and assume that for some t_0^* and T_0^* satisfying*

$$t_0^* > t_0 + 1, \qquad T_0^* \geq T_0 + 1,$$

$x|[t_0, t_0^*)$ is outside $y|[T_0, T_0^*)(x|(t_0, t_0^*]$ is outside $y|(T_0, T_0^*])$. If there exists $T_n > T_0^*$ for some $n \geq 1$ such that $y(T_n) = 0$ and y has n zeros on $(T_0, T_n]$, then there exists $t_n > t_0^*$ such that $x(t_n) = 0, x$ has n or $n+1$ zeros on $(t_0, t_n]$, and $x|[t_0, t_n + 1]$ is outside $y|[T_0, T_n + 1](x|(t_0, t_n + 1]$ is outside $y|(T_0, T_n + 1])$.

Moreover, if y has infinitely many zeros in $(T_0, +\infty)$, then x has infinitely many zeros in $(t_0, +\infty)$, and $x|[t_0, +\infty)$ is outside $y|[T_0, +\infty)(x|(t_0, +\infty)$ is outside $y|(T_0, +\infty))$.

Proof. The proof is omitted, because it is similar to the proof of [11, Lemma 3.1]. Here, we just make some comments on the proof. It should be pointed out that the key point in the proof of the Trajectory Crossing Lemma of [11] and [12] is that the orbits of x and y do not cross. The assumption in [11, Lemma 3.1], differing from ours, is that both x and y have infinitely many zeros. But this assumption is not needed in the proof of "noncrossing"; it is needed only in the proof of "outside." Since we assume that y has n zeros in $(T_0, T_n]$, it follows from "noncrossing" and Lemma 2.1 that there exists $t_n > t_0^*$ such that $x(t_n) = 0$ and x has n zeros on $(t_0, t_n]$. If $y(T_0) = 0$, there may be a $\bar{t}_0$ in (t_0, t_0^*) such that $x(\bar{t}_0) = 0$ and $(x(t), -x(t-1))$ is outside $(y(T_0), -y(T_0))$ for all t in $[t_0, \bar{t}_0]$. In this case, x has $n+1$ zeros in some interval $(t_0, t_n]$. By "noncrossing" again, one can see that $x|[t_0, t_n + 1]$, which has the same rotations around the origin as $y|[T_0, T_n + 1]$ does, is outside $y|[T_0, T_n + 1]$. Letting $n \to +\infty$, we know from the first part of the lemma that x has infinitely many zeros. Since the separations of zeros of x and y are greater than one by Lemma 2.1 and Definition 2.1, it follows that $T_n, t_n \to +\infty$ as $n \to +\infty$. Consequently, the second part of the lemma is true.

Note that the assertion in [11, Lemma 3.1] that $x|(t_0, +\infty)$ is outside $y|(T_0, +\infty)$ is not precise: it is not the case that if $x|[t_0, t_0^*)$ is outside $y|[T_0, T_0^*)$, then $x|(t_0, t_0^*)$ is outside $y|(T_0, t_0^*)$. A simple example is given for the case in which $y(t_0 - 1) = 0, x(t_0 - 1) > 0$, and $y(t_0) > x(t_0) > 0$. □

DEFINITION 2.2. *Suppose that x is a solution of* (2.1) *and has infinitely many zeros in $(t_0 - 1, +\infty)$. Assume $x_{t_0} \in C_*$. We say that x spirals outward (inward) in $(t_0, +\infty)$ if there exists some $t_1 > t_0$ such that $x|(t_1, +\infty)$ is outside $x|(t_0, +\infty)(x|(t_0, +\infty)$ is outside $x|(t_1, +\infty))$. We say x spirals toward an orbit γ in $\mathbb{R}^2$ if the point $(x(t), -x(t-1))$ in $\mathbb{R}^2$ approaches the orbit γ as $t \to +\infty$.*

For any slowly oscillating solution $x(t)$ of (2.1), one can see that $x(t_1 - 1) = 0$ for some t_1 implies that x has an extreme value at $t = t_1$. Therefore, by the definition of "outside," if $x|(t_1, +\infty)$ is outside $x|(t_0, +\infty)$ for some $t_1 > t_0$, then the orbit of x must make a complete rotation about the origin when t goes from t_0 to t_1. Let T be the map in Definition 2.1 making $x|(t_1, +\infty)$ outside $x|(t_0, +\infty)$. If $t_n = T^n(t_0)$ is the nth iteration of T at t_0 and $\tilde{T}^n$ is the restriction of T to the interval $(t_n, +\infty)$, then $x|(t_{n+1}, +\infty)$ is outside $x|(t_n, +\infty)$ with the corresponding mapping $\tilde{T}^n$ for $n = 0, 1, \ldots$. Thus, if x spirals outward, then the orbit of x in $\mathbb{R}^2$ rotates outward around

the origin infinitely. The discussion is similar if x spirals inward. Following the discussion above, we have a corollary of Lemma 2.2.

COROLLARY 2.3. *Suppose* (H1) *is satisfied. Let* x *be a solution of* (2.1) *on* $(t_0 - 1, +\infty)$. *Assume* $x_{t_0} \in C_*$ *and assume that there is a* $t_1 > t_0$ *such that* $x|(t_1, t_1 + 1]$ *is outside* $x|(t_0, t_0 + 1]$. *Then* $x|(t_1, +\infty)$ *is outside* $x|(t_0, +\infty)$ *and* x *spirals outward on* $(t_0, +\infty)$.

Proof. If x has n zeros in $(t_0, +\infty)$, following the discussion above, there exists at least one zero of x in the interval $(t_0, t_1]$. Therefore, x has at most $n - 1$ zeros in $(t_1, +\infty)$, contradicting Lemma 2.2. Thus x has infinitely many zeros. The proof is completed by the use of Lemma 2.2.

The following result presents a way to determine the existence of slowly oscillating periodic solutions. The result is implicitly given by [11] and its proof is omitted. □

THEOREM 2.4. *Suppose that* (H1) *is satisfied. Let* x *be a solution of* (2.1) *through* $(t_0, \varphi) \in \mathbb{R} \times C_*$. *If* x *is bounded and spirals outward in* $(t_0, +\infty)$, *then there exists a slowly oscillating periodic solution* $\tilde{x}$ *of* (2.1) *such that* x *spirals outward toward the orbit of* $\tilde{x}$. *If* x *spirals inward in* $(t_0, +\infty)$ *and the orbit of* x *is bounded away from the origin in* $\mathbb{R}^2$, *then there exists a slowly oscillating periodic solution* $\tilde{x}$ *of* (2.1) *such that* x *spirals inward toward the orbit of* $\tilde{x}$ *in* $\mathbb{R}^2$.

3. Multiexistence of slowly oscillating periodic solutions. In this section, we will present a checkable condition on the function f such that (2.1) has at least n distinct slowly oscillating periodic solutions.

THEOREM 3.1. *Suppose that* f *is an odd function and satisfies* (H1). *If there exist* $0 < a_1 < a_2 < \cdots < a_{2n}$ *such that, for* $k = 1, 2, \ldots, n$,

$$\int_0^{f(a_{2k-1})} f(x)\, dx > 2[f(a_{2k-1})]^2 \tag{3.1}$$

and

$$0 < f(a_{2k})/a_{2k} < 1, \tag{3.2}$$

then (2.1) *has at least* n *distinct slowly oscillating periodic solutions. In fact, there exist slowly oscillating solutions* $x^k(\cdot)$ *and slowly oscillating periodic solutions* $\bar{x}^k(\cdot)(k = 1, 2, \ldots, 2n)$ *such that*

(i) $x^{k+1}|(0, +\infty)$ *is outside* $x^k|(0, +\infty)$ *for* $k = 0, 1, 2, \ldots, 2n - 1$.

(ii) x^{2k-1} *spirals outward toward the orbit of* $\bar{x}^{2k-1}$, *and* x^{2k} *spirals inward toward the orbit of* $\bar{x}^{2k}$ *for* $k = 1, 2, \ldots, n$.

(iii) $\bar{x}^{2k-1}$ *and* $\bar{x}^{2k}$ *may be the same up to a time-shift*; $\bar{x}^{2k}$ *and* $\bar{x}^{2j}$ *are distinct for any* $k \neq j$.

If there exists an infinite sequence $\{a_n\}_{n=1}^{\infty}$ *such that* (3.1) *and* (3.2) *are satisfied for each* k, *then* (2.1) *has infinitely many slowly oscillating periodic solutions.*

To prove this theorem, we construct initial values $\varphi_1, \ldots, \varphi_{2n}$ in C_* such that, if $x^k = x(\cdot; \varphi_k)$ is a solution of (2.1) satisfying $x(\theta; \varphi_k) = \varphi_k(\theta)$ for $\theta \in [-1, 0]$, then $x^k(\cdot), k = 1, 2, \ldots, 2n$, satisfy (i) and (ii) of the theorem.

CLAIM A. Suppose that (H1) is satisfied and $\{a_k\}_1^{2n}$ is as in Theorem 3.1. Choose $\delta_{2k-1} > 0$ small enough that

$$\int_0^{f(a_{2k-1})(1-\delta_{2k-1})} f(x)\, dx > 2[f(a_{2k-1})]^2 \tag{3.3}$$

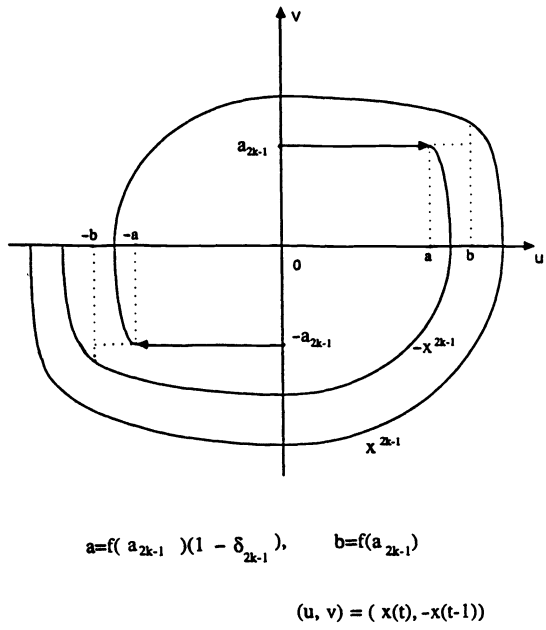


FIG. 1.

and

$$f(a_{2k-1})(1-\delta_{2k-1}) > a_{2k-1}. \tag{3.3)'}$$

If φ_{2k-1} in $C_*(k=1,2,\dots,n)$ is an arbitrary nonincreasing function satisfying

$$\varphi_{2k-1}(\theta)\begin{cases} = a_{2k-1}, & \theta \in [-1,-\delta_{2k-1}], \\ > 0, & \theta \in (-\delta_{2k-1},0), \\ = 0, & \theta = 0, \end{cases} \tag{3.4}$$

then the solution $x^{2k-1} = x(\cdot;\varphi_{2k-1})$ of (2.1) spirals outward in $(0,+\infty)$.

Proof. To prove the claim, we show the existence of the first two positive zeros, t_0 and t_2, of x^{2k-1} and then, prove that $x^{2k-1}|(t_2,t_2+1]$ is outside $x^{2k-1}|(0,1]$.

Note that from the inequality (3.1) and the assumption (H1), one can deduce that $f(a_{2k-1}) > a_{2k-1}$, and thus, there does exist $\delta_{2k-1} \in (0,1)$ such that (3.3) and (3.3)′ are satisfied. Let $x^{2k-1} = x(\cdot;\varphi_{2k-1})$ be the solution of (2.1) satisfying $x^{2k-1}(\theta) = \varphi_{2k-1}(\theta)$ for $\theta \in [-1,0]$. From (2.1), we have

$$x^{2k-1}(t) = -\int_0^t f(\varphi_{2k-1}(\theta-1))\,d\theta \quad \text{for } t \in [0,1].$$

Therefore,

$$x^{2k-1}(t) = -f(a_{2k-1})t \quad \text{for } t \in [0,1-\delta_{2k-1}] \tag{3.5}$$

and

$$-f(a_{2k-1}) < x^{2k-1}(t) < -f(a_{2k-1})(1-\delta_{2k-1}) \quad \text{for } t \in [1-\delta_{2k-1},1]. \tag{3.6}$$

Thus, the orbit of $x^{2k-1}|(0,1]$ in the (u,v) plane (that is, the $(x(t),-x(t-1))$ plane) connects the negative v axis to the negative u axis, and consists of two parts. The first part, $x^{2k-1}|(0,1-\delta_{2k-1}]$, lies in the third quadrant on the line $v=-a_{2k-1}$, while the other part, $x^{2k-1}|[1-\delta_{2k-1},1]$, is bounded by four lines, $v=-a_{2k-1}, v=0, u=-f(a_{2k-1})(1-\delta_{2k-1})$, and $u=-f(a_{2k-1})$. (See Fig. 1.) For t in $[1,2-\delta_{2k-1}]$, from (3.5) we know that

$$
\begin{aligned}
x^{2k-1}(t) &= x^{2k-1}(1) - \int_1^t f(-f(a_{2k-1})(s-1))\,ds \\
&= x^{2k-1}(1) + \int_0^{f(a_{2k-1})(t-1)} \frac{f(x)}{f(a_{2k-1})}\,dx,
\end{aligned} \tag{3.7}
$$

and therefore, for t in $[2-\delta_{2k-1}, 2]$,

$$
\begin{aligned}
x^{2k-1}(t) &\geq x^{2k-1}(2-\delta_{2k-1}) \\
&> -f(a_{2k-1}) + \int_0^{f(a_{2k-1})(1-\delta_{2k-1})} \frac{f(x)}{f(a_{2k-1})}\,dx \\
&> f(a_{2k-1}) > 0,
\end{aligned}
$$

by the first inequality of (3.6) and (3.3). Together with (3.5) and (3.6), the inequality above implies that the first positive zero t_0 of x^{2k-1} does exist and is in $(1, 2-\delta_{2k-1})$.

Now, we want to determine the location of the orbit of $x^{2k-1}|(t_0, t_0+1]$ in the (u, v) plane. Since $x^{2k-1}(t)$ is strictly increasing for $t \in (1, t_0+1)$ by (2.1), the inequality above also implies

$$
x^{2k-1}(t) > f(a_{2k-1}) \quad \text{for } t \in [2-\delta_{2k-1}, t_0+1]. \tag{3.8}
$$

If $t_1 = a_{2k-1}/f(a_{2k-1})$, then t_1 is in $(0, 1-\delta_{2k-1})$ by the choice of δ_{2k-1}. We want to show $t_0 - 1 > t_1$. Using the second inequality of (3.6) and the monotonicity of f we deduce from (3.7) that, for t in $[1, t_1+1] \subset [1, 2-\delta_{2k-1}]$,

$$
\begin{aligned}
x^{2k-1}(t) &\leq x^{2k-1}(1) + \int_0^{f(a_{2k-1})t_1} \frac{f(x)}{f(a_{2k-1})}\,dx \\
&< -f(a_{2k-1})(1-\delta_{2k-1}) + \int_0^{a_{2k-1}} \frac{f(x)}{f(a_{2k-1})}\,dx \\
&< -f(a_{2k-1})(1-\delta_{2k-1}) + a_{2k-1} < 0,
\end{aligned}
$$

by the choice of δ_{2k-1}. Together with (3.5) and (3.6), the inequality above implies $x^{2k-1}(t) < 0$ for all $t \in (0, t_1+1]$ and thus, $t_1 + 1 < t_0$ or $t_0 - 1 > t_1$, where t_0 in $(1, 2-\delta_{2k-1})$ is the first positive zero of x^{2k-1}. Consequently, it follows from (3.5) and the definition of t_1 that

$$
x^{2k-1}(t) < x^{2k-1}(t_1) = -a_{2k-1} \tag{3.8}'
$$

for t in $[t_0-1, 1-\delta_{2k-1}] \subset (t_1, 1-\delta_{2k-1})$. The discussion above shows that the orbit of $x^{2k-1}|(t_0, t_0+1]$ in the (u, v) plane connects the positive v axis to the positive u axis, and consists of two parts. The first part, $x^{2k-1}|(t_0, 2-\delta_{2k-1}]$, lies in the first quadrant above the line $v = a_{2k-1}$ by (3.8)′, while the other part, $x^{2k-1}|[2-\delta_{2k-1}, t_0+1]$, is on the right side of the line $u = f(a_{2k-1})$ by (3.8). (See Fig. 1.)

Since f is an odd function, $\tilde{x}^{2k-1}(t) = -x(t; \varphi_{2k-1}), t \in [-1, +\infty)$, is a solution of (2.1) satisfying $\tilde{x}^{2k-1}(\theta) = -\varphi_{2k-1}(\theta)$ for θ in $[-1, 0]$. From the discussions above one can see that $x^{2k-1}|(t_0, t_0+1]$ is outside $\tilde{x}^{2k-1}|(0, 1]$ and $\tilde{x}|(t_0, t_0+1]$ is outside $x^{2k-1}|(0, 1]$. (See Fig. 1.) The Trajectory Crossing Lemma implies that there exists $t_2 > t_0 + 1$ such that $x^{2k-1}(t_2) = 0$ and $x^{2k-1}|(t_0, t_2+1]$ is outside $\tilde{x}|(0, t_0+1]$.

It is easy to see that the counterpart of $x^{2k-1}|(t_2, t_2+1]$ is $\tilde{x}^{2k-1}|(t_0, t_0+1]$—both orbits connect the negative v axis to the negative u axis—and therefore, the former is outside the latter. Consequently, $x^{2k-1}|(t_2, t_2+1]$ is outside $x^{2k-1}|(0,1]$, since $\tilde{x}^{2k-1}|(t_0, t_0+1]$ is outside the latter, as we mentioned above. The proof is completed by Corollary 2.3. □

CLAIM B. Suppose that (H1) is satisfied and $\{a_k\}_1^{2n}$ is as in Theorem 3.1. Choose $\delta_{2k} > 0$ small enough that

$$f(a_{2k})(1-\delta_{2k}) > \max\{f(a_{2k-1}), f(f(a_{2k}))\}. \tag{3.9}$$

Let $\varphi_{2k} (k = 1, 2, \ldots, n)$ in C_* be an arbitrary nonincreasing function satisfying

$$\varphi_{2k}(\theta) \begin{cases} = a_{2k}, & \theta \in [-1, -\delta_{2k}], \\ > 0, & \theta \in (-\delta_{2k}, 0), \\ = 0, & \theta = 0. \end{cases} \tag{3.10}$$

If $x^{2k} = x(\cdot; \varphi_{2k})$ is the solution of (2.1) and x^{2k-1} is as in Claim A, then $x^{2k}|(0, +\infty)$ is outside $x^{2k-1}|(0, +\infty)$ and x^{2k} spirals inward in $(0, +\infty)$.

Proof. By the assumptions we know that $a_{2k} > a_{2k-1} > 0$ and $a_{2k} > f(a_{2k}) > 0$; thus it follows from the monotonicity of f that there exists $\delta_{2k} > 0$ satisfying (3.9). Suppose $x^{2k} = x(\cdot; \varphi_{2k})$ is the solution of (2.1) satisfying $x^{2k}(\theta) = \varphi_{2k}(\theta)$ for θ in $[-1, 0]$. From (2.1), we have

$$x^{2k} = -\int_0^t f(\varphi_{2k}(\theta - 1))\, d\theta \quad \text{for } t \in [0, 1].$$

Therefore,

$$x^{2k}(t) = -f(a_{2k})t \quad \text{for } t \in [0, 1-\delta_{2k}] \tag{3.11}$$

and

$$-f(a_{2k}) < x^{2k}(t) < -f(a_{2k})(1-\delta_{2k}) \quad \text{for } t \in (1-\delta_{2k}, 1]. \tag{3.12}$$

Thus, the orbit of $x^{2k}|(0,1]$ in the (u, v) plane connects the negative v axis to the negative u axis, and consists of two parts. The first part, $x^{2k}|(0, 1-\delta_{2k}]$, lies in the third quadrant on the line $v = -a_{2k}$, while the other part, $x^{2k}|(1-\delta_{2k}, 1]$, is bounded by the four lines $v = -a_{2k}, v = 0, u = -f(a_{2k})$, and $u = -f(a_{2k})(1-\delta_{2k})$. Since $a_{2k} > a_{2k-1}$ and $f(a_{2k})(1-\delta_{2k}) > f(a_{2k-1})$, from the discussion right after (3.6), we know that the first and the second parts of $x^{2k}|(0,1]$ are below and on the right side of the orbit $x^{2k-1}|(0,1]$, respectively. Thus, $x^{2k}|(0,1]$ is outside $x^{2k-1}|(0,1]$. Therefore, by Claim A and Lemma 2.2, $x^{2k}|(0, +\infty)$ is outside $x^{2k-1}|(0, +\infty)$ and x^{2k} has infinitely many zeros in $(0, +\infty)$.

Let $t = t_0$ be the first positive zero of x^{2k}. One can see that $t_0 > 1$, because x^{2k} is a slowly oscillating solution and $x^{2k}(0) = 0$. Since $x^{2k}(t)$ is increasing in $(1, t_0 + 1)$, we know that

$$-f(a_{2k}) < x^{2k}(t) < 0 \tag{3.13}$$

for $t \in [t_0 - 1, t_0) \subset [0, t_0)$, and

$$\begin{aligned} x^{2k}(t) &= -\int_{t_0}^t f(x^{2k}(s-1))\, ds \\ &< f(f(a_{2k}))(t - t_0) \le f(f(a_{2k})) \quad \text{for } t \in (t_0, t_0 + 1]. \end{aligned} \tag{3.14}$$

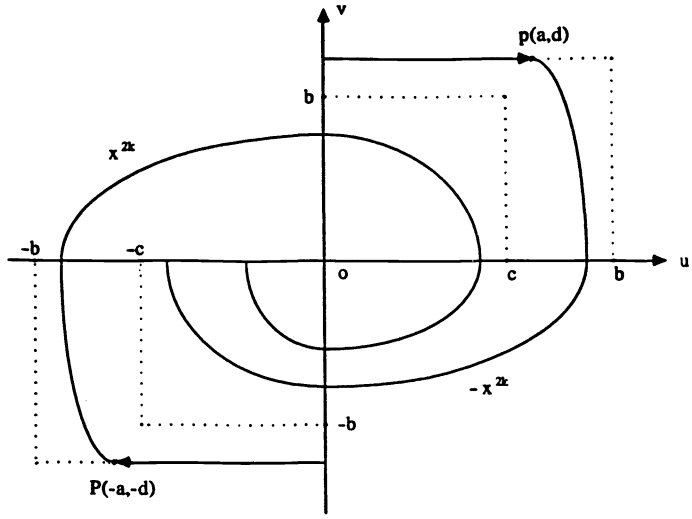


FIG. 2.

Therefore, $x^{2k}|(t_0, t_0+1]$ connects the positive v axis to the positive u axis and is bounded by the four lines $v = f(a_{2k}), v = 0, u = f(f(a_{2k}))$, and $u = 0$.

By the oddness of $f, \tilde{x}^{2k}(t) = -x^{2k}(t)$ is the solution of (2.1) satisfying $\tilde{x}^{2k}(\theta) = -\varphi_{2k}(\theta)$ for θ in $[-1, 0]$. Through the discussion above, we can see that $\tilde{x}^{2k}|(0,1]$ is outside $x^{2k}|(t_0, t_0+1]$ and $x^{2k}|(0,1]$ is outside $\tilde{x}^{2k}|(t_0, t_0+1]$, since $a_{2k} > f(a_{2k})$ and $f(a_{2k})(1-\delta_{2k}) > f(f(a_{2k}))$. (See Fig. 2.) Therefore, by Lemma 2.2, it is easy to see that $\tilde{x}^{2k}|(0, t_0+1]$ is outside $x^{2k}|(t_0, t_2+1]$, where $t = t_2$ is the second zero of x^{2k} for $t > 0$. One can see that the counterpart of $\tilde{x}^{2k}|(t_0, t_0+1]$ is $x^{2k}|(t_2, t_2+1]$—both orbits connect the negative v axis to the negative u axis—and therefore, the former is outside the latter. (See Fig. 2.) Consequently, $x^{2k}|(0,1]$ is outside $x^{2k}|(t_2, t_2+1]$ because $x^{2k}|(0,1]$ is outside $\tilde{x}^{2k}|(t_0, t_0+1]$. By Lemma 2.2, $x^{2k}|(0, +\infty)$ is outside $x^{2k}|(t_2, +\infty)$ since we have shown that x^{2k} has infinitely many zeros in $(0, +\infty)$. This means x^{2k} spirals inward in $(0, +\infty)$ by the definition. Claim B is proven. □

Proof of Theorem 3.1. If in Claim A we choose $\delta_{2k-1} > 0 (k = 2, 3, \ldots, n)$ so small that

$$f(a_{2k-1})(1-\delta_{2k-1}) > f(a_{2k-2})$$

is also satisfied, then similar to the discussions above one can see that $x^{2k-1}|(0,1]$ is outside $x^{2k-2}|(0,1]$ (since both of them connect the negative v axis to the negative u axis); $x^{2k-1}|(0,1]$ is in the exterior of the rectangle bounded by $v = -a_{2k-1}, v = 0, u = 0$, and $u = -f(a_{2k-1})$; and $x^{2k-2}|(0,1]$ is in the interior of that rectangle. Therefore, we have constructed solutions $x^k(\cdot)$ for $k = 1, 2, \ldots, 2n$ satisfying the following.

(i)′ $x^{k+1}|(0, +\infty)$ is outside $x^k|(0, +\infty)$ for $k = 1, 2, \ldots, 2n-1$,

(ii)′ x^{2k-1} spirals outward and x^{2k} spirals inward for $k = 1, 2, \ldots, n$.

The statement that x^{2k-1} spirals outward implies that the orbit of x^{2k-1} is bounded away from the origin of $\mathbb{R}^2$, and the statement that x^{2k} spirals inward implies that x^{2k} is bounded. On the other hand, by the property (i)′ above, the orbit of x^{2k} keeps

x^{2k-1} bounded and the orbit of x^{2k-1} keeps the orbit of x^{2k} bounded away from the origin. Therefore, by Theorem 2.4, there exist slowly oscillating periodic solutions $\bar{x}^{2k}$ and $\bar{x}^{2k-1}$ of (2.1) such that x^{2k-1} spirals outward toward the orbit of $\bar{x}^{2k-1}$ in $\mathbb{R}^2$ and x^{2k} spirals inward toward the orbit of $\bar{x}^{2k}$ in $\mathbb{R}^2$. It is obvious that $\bar{x}^{2k}$ and $\bar{x}^{2j}$ are distinct for any $k \neq j$. The proof is completed. □

4. An example. In this section, an equation is constructed which satisfies the conditions in Theorem 3.1, and which therefore has infinitely many slowly oscillating periodic solutions. Consider the equation

$$\dot{x}(t) = -f(x(t-1)), \tag{4.1}$$

where $f(x) = (b + \frac{3}{4})x + bx \sin[\ln(1 + |x|^{2/(3b)})]$.

PROPOSITION 4.1. *If $b \geq \frac{13}{4}$, then (4.1) has infinitely many slowly oscillating periodic solutions.*

Proof. It is obvious that $f(0) = 0$ and f is an odd function.

$$\begin{aligned} f'(x) &= \left(b + \frac{3}{4}\right) + b \sin[\ln(1 + |x|^{2/(3b)})] \\ &\quad + \frac{\frac{2}{3}|x|^{2/(3b)}}{1 + |x|^{2/(3b)}} \cos[\ln(1 + |x|^{2/(3b)})] \\ &\geq \frac{3}{4} - \frac{\frac{2}{3}|x|^{2/(3b)}}{1 + |x|^{2/(3b)}} > \frac{1}{12} \end{aligned}$$

for all $x \in \mathbb{R}$. Thus, if $x = f^{-1}(y)$ is the inverse of $y = f(x)$ and $\{a_k\}$ is defined by

$$a_{2k} = f^{-1}((e^{(2k-\frac{1}{2})\pi} - 1)^{3b/2}), \qquad a_{2k-1} = f^{-1}((e^{(2k-1)\pi} - 1)^{3b/2}) \tag{4.2}$$

for $k = 1, 2, 3, \ldots$, then $0 < a_1 < a_2 < a_3 < \cdots < a_{2k-1} < a_{2k} < \cdots$, and

$$f(a_{2k}) = \left(b + \frac{3}{4}\right) a_{2k} - (b) a_{2k} < a_{2k}.$$

Consequently,

$$\begin{aligned} \int_0^{f(a_{2k-1})} f(x)\, dx &= \frac{1}{2}\left(b + \frac{3}{4}\right)[f(a_{2k-1})]^2 + \int_0^{f(a_{2k-1})} bx \sin[\ln(1 + |x|^{2/(3b)})] dx \\ &= \frac{1}{2}\left(b + \frac{3}{4}\right)[f(a_{2k-1})]^2 + \int_0^{(2k-1)\pi} \frac{3b^2}{2}(e^s - 1)^{3b-1} e^s \sin s\, ds, \end{aligned}$$

where $s = \ln(1 + |x|^{2/(3b)})$ or $x = (e^s - 1)^{3b/2}$. If $h(s) = (e^s - 1)^{3b-1} e^s$, then $h(s)$ is positive and increasing for $s > 0$. Therefore,

$$\int_{(2j-1)\pi}^{(2j+1)\pi} \frac{3b^2}{2} h(s) \sin s\, ds > 0 \quad \text{for } j \geq 1,$$

and

$$\int_0^{f(a_{2k-1})} f(x)\, dx > \frac{1}{2}\left(b + \frac{3}{4}\right) f^2(a_{2k-1}) \geq 2 f^2(a_{2k-1}).$$

Thus, the function f satisfies the condition in Theorem 3.1 with $\{a_k\}_{k=1}^{\infty}$ defined by (4.2). Consequently, (4.1) has infinitely many slowly oscillating periodic solutions. □

Remark. For the more general function $f(x) = b_1 x + b_2 x \sin[\ln(1 + |x|^{\mu})]$, if $b_1 \geq 4, b_1 > b_2 > b_1 - 1$, and $0 < \mu < (b_1 - b_2)/b_2$, then using the technique of the preceding theorem one can show that this function f satisfies the conditions in Theorem 3.1 with $\{a_k\}_{k=1}^{\infty}$ defined by

$$a_{2k} = f^{-1}((e^{(2k-\frac{1}{2})\pi} - 1)^{1/\mu}), \qquad a_{2k-1} = f^{-1}((e^{(2k-1)\pi} - 1)^{1/\mu}). \tag{4.3}$$

Therefore, the equation (4.1) with this function f has infinitely many slowly oscillating periodic solutions.

REFERENCES

[1] R. Bellman and K. L. Cooke, *Differential-Difference Equations,* Academic Press, New York, 1963.

[2] S.-N. Chow, *Existence of periodic solutions of autonomous functional differential equation,* J. Differential Equations, 17 (1974), pp. 365–378.

[3] S.-N. Chow and J. Mallet-Paret, *Integral averaging and bifurcation,* J. Differential Equations, 26 (1977), pp. 112–159.

[4] ———, *The Fuller index and global Hopf bifurcation,* J. Differential Equations, 29 (1978), pp. 66–85.

[5] W. J. Cunningham, *A nonlinear differential-difference equation of growth,* Proc. Nat. Acad. Sci. USA, 40 (1974), pp. 709–713.

[6] R. B. Grafton, *A periodicity theorem for autonomous functional differential equations,* J. Differential Equations, 6 (1969), pp. 87–109.

[7] K. P. Hadder and J. Tomiuk, *Periodic solutions of difference-differential equations,* Arch. Rational Mech. Anal., 65 (1977), pp. 82–95.

[8] J. K. Hale, *Theory of Functional Differential Equations,* Springer-Verlag, Berlin, 1977.

[9] G. S. Jones, *The existence of periodic solutions of* $f'(x) = \alpha f(x)\{1 + f(x)\}$, J. Math. Anal. Appl., 5 (1962), pp. 535–450.

[10] J. L. Kaplan and J. A. Yorke, *Ordinary differential equations which yield periodic solutions of differential delay equations,* J. Math. Anal. Appl., 48 (1974), pp. 317–324.

[11] ———, *On the stability of a periodic solution of a differential delay equation,* SIAM J. Math. Anal., 6 (1975), pp. 268–282.

[12] ———, *On the nonlinear differential delay equation* $\dot{x}(t) = -f(x(t), x(t-1))$, J. Differential Equations, 23 (1977), pp. 293–314.

[13] R. D. Nussbaum, *Periodic solutions of some nonlinear autonomous functional differential equations,* Amer. Math. Pura. Appl., 101 (1974), pp. 263–306.

[14] ———, *Periodic solutions of some nonlinear autonomous functional differential equations* II, J. Differential Equations, 14 (1973), pp. 360–394.

[15] ———, *A global bifurcation theorem with applications to functional differential equations,* J. Funct. Anal., 19 (1975), pp. 319–338.

[16] ———, *Periodic solutions of nonlinear autonomous functional differential equations,* Lecture Notes in Math. 730, Springer-Verlag, Berlin, 1979, pp. 283–325.

[17] ———, *Uniqueness and nonuniqueness for periodic solutions of* $x'(t) = -g(x(t-1))$, J. Differential Equations, 34 (1974), pp. 25–54.

[18] ———, *The range of periods of periodic solutions of* $x'(t) = -\alpha f(x(t-1))$, J. Math. Anal. Appl., 58 (1977), pp. 280–292.

[19] H. O. Walther, *A theorem on the amplitudes of periodic solution of differential delay equations with applications of bifurcation,* J. Differential Equations, 29 (1978), pp. 396–404.

[20] E. M. Wright, *A non-linear differential equation,* J. Reine Angew. Math., 194 (1955), pp. 66–87.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 446–467, March 1995

011

ON RECURRENCE RELATIONS FOR SOBOLEV ORTHOGONAL POLYNOMIALS *

W. D. EVANS†, LANCE L. LITTLEJOHN‡, FRANCISCO MARCELLAN§, CLEMENS MARKETT¶, AND ANDRE RONVEAUX‖

Abstract. This paper discusses recurrence relations for sequences of polynomials which are orthogonal with respect to the Sobolev inner product defined on the set of polynomials $\mathcal{P}$ by

$$(p,q)_W = \sum_{k=0}^{N} \int_{\mathbb{R}} p^{(k)}(x)\bar{q}^{(k)}(x)\, d\mu_k(x) \qquad (p,q \in \mathcal{P})$$

for some integer $N \geq 1$, where each $\mu_k, 0 \leq k \leq N$, is a positive Borel measure. It is proven that there exists a real-valued polynomial $h : \mathbb{R} \to \mathbb{R}$ satisfying

$$(*) \qquad (hp,q)_W = (p,hq)_W \qquad (p,q \in \mathcal{P})$$

if and only if each of the measures $\mu_k, 1 \leq k \leq N$, is purely atomic with a finite number of mass points. In addition it is proven that R_j, the set of real roots of $d^j h/dx^j, (1 \leq j \leq N)$, is nonempty and that supp $(\mu_k) \subset \cap_{i=1}^{k} R_i$. It is also shown that if h satisfies the condition $(*)$, then the polynomials orthogonal with respect to the inner product $(\cdot,\cdot)_W$ will satisfy a recurrence relation of order $2m+1$, where $m = \deg(h)$. Furthermore, an algorithm is given to construct a polynomial H of minimal positive degree for which the above properties hold. Several examples will be discussed to illustrate the theory. Lastly it is shown, under certain circumstances, when these orthogonal polynomials will satisfy second-order linear differential equations.

Key words. Sobolev orthogonal polynomials, Borel measures, Dirac point mass measures, recurrence relations, second-order differential equations, structural relations

AMS subject classifications. 33A65, 28A25

1. Introduction.

It is well known (i.e., see [3, pp. 21–22]) that every sequence of polynomials $\{\phi_n(x)\}_{n=0}^{\infty}$ orthogonal with respect to an inner product of the form

$$(1.1) \qquad (p,q)_\mu := \int_{\mathbb{R}} p(x)\bar{q}(x)\, d\mu(x),$$

where μ is a signed measure on the real line $\mathbb{R}$, satisfies a three-term recurrence relation of the form

$$(1.2) \qquad \begin{aligned} \phi_{n+1}(x) &= (a_n x + b_n)\phi_n(x) - c_n\phi_{n-1}(x) \qquad (n \geq 0), \\ \phi_0(x) &= c_0; \phi_{-1}(x) \equiv 0, \end{aligned}$$

* Received by the editors March 2, 1992; accepted for publication (in revised form) October 6, 1993.

† School of Mathematics, University of Wales, Senghennydd Road, Cardiff, Wales, CF2 4AG, United Kingdom.

‡ Department of Mathematics, Utah State University, Logan, Utah, 84322-3900.

§ Departamento de Matemática, Universidad Carlos III de Madrid, Avda. Mediterraneo, S/N, 28913-Leganés, Madrid, Spain. This author's research was partially supported by the Comisión Interministerial de Ciencia y Tecnologia of Spain under grant PB89-0181-CO2-01.

¶ Lehrstuhl A für Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, D-5100 Aachen, Germany.

‖ Laboratoire de Physique Mathématique, Facultés Universitaires Notre-Dame de la Paix, Rue de Bruxelles, 61, 5000 Namur, Belgium. This author's research was partially supported by the Fonds National de la Recherche Scientifique of Belgium.

where $\{a_n\}, \{b_n\}, \{c_n\} \subset \mathbb{R}$ and $a_n c_n \neq 0, n = 0, 1, 2, \ldots$. The key step in establishing (1.2) involves dealing with the obvious identity

$$(xp(x), q(x))_\mu = (p(x), xq(x))_\mu \tag{1.3}$$

for all polynomials p, q.

During the last few years, there have been several papers written about polynomials orthogonal with respect to Sobolev inner products. Some of these inner products are of the form

$$(p, q) := \int_I p(x)\bar{q}(x)\, d\mu(x) + \sum_{r=1}^{r_1} M_r p(c_r)\bar{q}(c_r) + \sum_{r=1}^{r_2} N_r p'(d_r)\bar{q}'(d_r), \tag{1.4}$$

where I is some interval on the real line, $M_r \geq 0, N_r \geq 0$, and c_r, d_r are fixed points (not necessarily in I). For example, see the contributions [1], [2], [12], [14], and the references therein. In each of the examples in these papers, the authors exhibit a polynomial $h(x)$ of degree $m \geq 1$ such that

$$(hp, q) = (p, hq)$$

for all polynomials p and q, and from this deduce that the corresponding orthogonal polynomials $\{\phi_n(x)\}$ satisfy a recurrence relation of the form

$$h(x)\phi_n(x) = \sum_{k=n-m}^{n+m} b_{n,k}\phi_k(x).$$

For Sobolev inner products involving derivatives of higher order, see [12].

We remark that all of this work has been influenced and popularized, in one way or another, by the paper of A. M. Krall [10], which has led to attaching mass points to the absolutely continuous measures associated with the classical orthogonal polynomials in order to obtain new orthogonal polynomials satisfying certain fourth-order differential equations.

Note that, in (1.4), if $N_r > 0$ for some $r, 1 \leq r \leq r_2$, it will *not* be the case that

$$(xp(x), q(x)) = (p(x), xq(x))$$

for all polynomials p, q. However, the authors in the above-mentioned papers exhibit polynomials $h : \mathbb{R} \to \mathbb{R}$ of degree $m = m(r_2) \geq 2$ such that

$$(h(x)p(x), q(x)) = (p(x), h(x)q(x))$$

for all polynomials p, q and, from this, deduce that $\{\phi_n(x)\}$ satisfies a $(2m+1)$-term recurrence relation of the form

$$h(x)\phi_n(x) = \sum_{k=n-m}^{n+m} b_{n,k}\phi_k(x) \qquad (n = m, m+1, \ldots).$$

In this paper, we shall be concerned with the weighted Sobolev inner product

$$(p, q)_W := \sum_{k=0}^{N} \int_{\mathbb{R}} p^{(k)}(x)\bar{q}^{(k)}(x)\, d\mu_k(x) \tag{1.5}$$

on the set

$$\mathcal{P} := \left\{ p(x) = \sum_{k=0}^{n} a_k x^k | a_k \in \mathbb{C}, n \in \mathbb{N}_0 \right\}$$

of all polynomials $p : \mathbb{R} \to \mathbb{C}$, where each $\mu_k, k = 0, 1, \ldots, N$, is a positive Borel measure on the Borel subsets $\mathcal{B}(\mathbb{R})$ of the real line $\mathbb{R}$, and $N \geq 1$ is a fixed integer. We shall also assume that the moments $\{c_{n,k}\}_{n=0}^{\infty}$ of $\mu_k, 0 \leq k \leq N$, defined by

$$c_{n,k} := \int_{\mathbb{R}} x^n \, d\mu_k \qquad (n = 0, 1, 2, \ldots; k = 0, 1, \ldots, N),$$

exist and are finite. Notice that the inner product in (1.5) includes every inner product of the form (1.4). We shall also find it necessary in certain situations to let $\mathcal{H}$ denote a Hilbert space generated by the inner product $(\cdot, \cdot)_W$ satisfying the inclusion $\mathcal{P} \subset \mathcal{H}$. It is not assumed that $\mathcal{P}$ is dense in $\mathcal{H}$. We note that an example of $\mathcal{H}$ could be the classical weighted Sobolev space $W_2^N(\mathbb{R}, d\mu_0, \ldots, d\mu_N)$ (see [11]).

Before stating the main theorem of this paper, we note that we shall use the notation δ_{x_0} to denote the Dirac point mass measure defined on the Borel subsets of $\mathbb{R}$ by

$$\delta_{x_0}(B) = \begin{cases} 1 & \text{if } x_0 \in B \\ 0 & \text{if } x_0 \notin B \end{cases} \qquad (B \in \mathcal{B}(\mathbb{R})).$$

Moreover, as is customary, we shall write $d(\delta_{x_0}) = \delta(x - x_0)dx$, where $\delta(x - x_0)$ is the so-called Dirac delta distribution. We list other notation that the reader will frequently encounter in this paper:

$\mathbb{N}_0$ denotes the set of nonnegative natural numbers;
$\mathbb{C}$ denotes the set of complex numbers;
$\bar{z}$ denotes the complex conjugate of the complex number z;
$f^{(k)}(x)$ denotes the kth derivative of the function f with respect to x;
supp (μ) denotes the support of the positive measure μ.

In §3 we shall prove the following theorem.

THEOREM 1.

(i) *Suppose there exists a polynomial $h : \mathbb{R} \to \mathbb{R}$ of degree ≥ 1 satisfying*

$$(hp, q)_W = (p, hq)_W \qquad (p, q \in \mathcal{P}), \tag{1.6}$$

where $(\cdot, \cdot)_W$ is the inner product defined in (1.5). *Then the measures $\mu_k, 1 \leq k \leq N$ are necessarily of the form*

$$\mu_k = \sum_{j=1}^{P(k)} \alpha_{k,j} \delta_{x_{k,j}} \tag{1.7}$$

for some positive integer $P(k)$, where

(a) $\alpha_{k,j} \geq 0, j = 1, \ldots, P(k), k = 1, \ldots, N$;

(b) $\{x_{k,j} | j = 1, \ldots, P(k)\} := R_k \neq \emptyset$ *are the distinct real roots of $h^{(k)}(x), 1 \leq k \leq N$;*

(c) supp $(\mu_k) \subset \cap_{j=1}^{k} R_j \neq \emptyset, k = 1, \ldots, N$.

Moreover, the degree of h is at least $N + 1$ and

(d) *there exists a unique (up to a nonzero, real constant multiple) polynomial $H : \mathbb{R} \to \mathbb{R}$ of minimal degree $m(H)$ $(\geq N + 1)$ satisfying $H(0) = 0$ and*

$$(Hp, q)_W = (p, Hq)_W \qquad (p, q \in \mathcal{P});$$

(e) *the sequence of polynomials* $\{\phi_n(x)\}$ *that are orthogonal with respect to the inner product* $(\cdot,\cdot)_W$ *will satisfy a* $(2m(H)+1)$*-term recurrence relation*

$$H(x)\phi_n(x) = \sum_{k=n-m(H)}^{n+m(H)} b_{n,k}\phi_k(x) \qquad (n \geq m(H)) \tag{1.8}$$

for some real numbers $b_{n,k}$. *Furthermore, the* length (*see Remark* 8 *in* §3, *below*) *of this recurrence relation is at least* $2N+3$.

(ii) *Suppose the measures* $\mu_k, 1 \leq k \leq N$, *in the inner product* (1.5) *are given by*

$$\mu_k = \sum_{j=1}^{Q(k)} \beta_{kj}\delta_{y_{kj}}$$

for some positive integer $Q(k)$ *with* $\operatorname{supp}(\mu_k) = \{y_{k1}, \ldots, y_{k,Q(k)}\}$ (*i.e.,* $\beta_{kj} > 0$ *for* $j = 1, \ldots, Q(k)$). *Then there exists a unique* (*up to a nonzero, real constant multiple*) *polynomial* $H : \mathbb{R} \to \mathbb{R}$ *of minimal degree* $m(H) \geq N+1$ *satisfying the conditions of* (d) *and* (e) *above. Moreover, if* R_k *denotes the set of real roots of* $H^{(k)}(x), 1 \leq k \leq N$, *then* (c) *is valid.* □

Remark 1. The fact that $h^{(k)}(x)$ has at least one real root $(1 \leq k \leq N)$ may seem somewhat surprising upon first glance, but it is a consequence of (1.6) (see Theorem 4 below) and not a condition that is assumed.

Remark 2. Under the conditions of Theorem 1, we see that the measures $\mu_k, 1 \leq k \leq N$, are *purely atomic.* No restriction is placed on the leading measure μ_0; see §5 of this paper for further discussion of this observation.

Remark 3. The Gram–Schmidt orthogonalization process (see [3, pp. 13–14]) may be used to assert the existence of a sequence $\{\phi_n(x)\}$ of *real* polynomials which are orthogonal with respect to the inner product $(\cdot,\cdot)_W$, defined in (1.5). Of course, it may be difficult in practice to explicitly compute each $\phi_n(x)$ by this method. If this is the case and if the recurrence coefficients b_{nk} can be efficiently calculated, then the recurrence relation (1.8) affords an alternative way of computing each $\phi_n(x)$.

The reader should also consult the contribution of Iserles et al. [8] for another approach to effectively finding each $\phi_n(x)$. They consider the inner product (1.5) in the case of $N = 1$ and when the positive Borel measure μ_1 has infinite support. In this case, the polynomials $\{\phi_n(x)\}$ will not necessarily satisfy a recurrence relation of the form (1.8). Consequently, the *computation* of these polynomials will, in general, be quite difficult. However, in [8], the authors give conditions on when these polynomials can be efficiently computed in terms of the orthogonal polynomials associated with the leading measure μ_0.

Throughout this paper we shall speak of *the* orthogonal polynomial sequence $\{\phi_n(x)\}$; of course, each $\phi_n(x)$ is only uniquely determined up to an arbitrary nonzero constant multiple.

Remark 4. As mentioned above, the key step in establishing a three-term recurrence relation for an orthogonal polynomial sequence in a space generated by the inner product (1.1) is in recognizing the identity (1.3). It is interesting to point out that condition (1.3) may be stated in terms of Hermitian operators. Indeed, if $\mathcal{H}$ denotes the Hilbert space generated by $(\cdot,\cdot)_\mu$, then (1.3) is equivalent to the operator $T : \mathcal{P} \to \mathcal{H}$ being Hermitian, where T is defined by

$$T(p(x)) = xp(x), \qquad p \in \mathcal{D}(T) = \mathcal{P}.$$

Moreover, if $\mathcal{P}$ is dense in $\mathcal{H}$, then (1.3) is equivalent to the operator T being symmetric. See Remark 10, below, for further discussion.

At this point, we mention the important work of Duran [4], which we just recently obtained. For each $N \geq 1$, he characterizes those inner products $(\cdot, \cdot)$ on the space $\mathcal{P}$ of polynomials for which the operator $S_N : \mathcal{P} \to \mathcal{P}$ defined by

$$S_N(p(x)) = x^N p(x), \qquad p \in \mathcal{D}(S_N) = \mathcal{P}$$

is Hermitian. From this characterization, he obtains inner products more general than our inner product in (1.5). In §3 of his paper, and by considering an approach different from ours, he does consider the inner product (1.5). Although he shows that the measures $\mu_k, 1 \leq k \leq N$, are discrete in this case, his characterization is not as explicit as our Theorem 1.

Remark 5. In this paper, we shall only consider *real-valued* polynomials h of degree ≥ 1 satisfying (1.6). Further work needs to be done in the case of complex-valued polynomials h satisfying (1.6). The interesting case seems to be when h is real-valued; indeed, this seems to be the only case considered in the literature.

Remark 6. We could have defined our inner product $(\cdot, \cdot)_W$ in (1.5) to be

$$(p, q)_W = \sum_{k=0}^{N} \int_{I_k} p^{(k)}(x) \bar{q}^{(k)}(x)\, d\mu_k. \tag{1.9}$$

where I_k is a real interval $(0 \leq k \leq N)$. A particularly interesting case would arise when $I_k \cap I_j = \emptyset, k, j = 0, \ldots, N$. However, since each μ_k can be extended in an obvious way to a Borel measure on all of $\mathcal{B}(\mathbb{R})$, we see that the inner product (1.9) is a special case of (1.5). We remark, however, that there is at least one *disadvantage* of using (1.5) over (1.9): the inner product in (1.9) can be more descriptive. For example, suppose $\{\phi_n(x)\}$ is an orthogonal polynomial sequence with respect to the inner product

$$(p, q)_I := \int_I p(x)\bar{q}(x)\, d\mu(x),$$

where μ is a positive measure and I is some interval on the real line. The classical theory of orthogonal polynomials dictates that these polynomials will have their roots in the interior of I. This may not be so clear if we write the above inner product as

$$\int_{\mathbb{R}} p(x)\bar{q}(x)\, d\mu(x).$$

Remark 7. Before proceeding with the rest of the paper, we point out that if an orthogonal polynomial sequence (OPS) $\{\phi_n(x)\}$ is orthogonal with respect to the inner product

$$(p, q)_W := \sum_{k=0}^{N} \int_{\mathbb{R}} p^{(k)}(x) \bar{q}^{(k)}(x)\, d\mu_k(x)$$

as well as orthogonal with respect to the inner product defined from the leading measure μ_0, i.e.

$$(p, q)_{\mu_0} := \int_{\mathbb{R}} p(x)\bar{q}(x)\, d\mu_0(x),$$

then

(i) $\{\phi_n(x)\}$ will satisfy a three-term recurrence relation, and

(ii) $\mu_k, 1 \leq k \leq N$, does not necessarily have the form (1.7), and

(iii) there does not necessarily exist a real-valued polynomial $h(x)$ of degree ≥ 1 such that (1.6) holds.

Indeed, some examples are the classical orthogonal polynomials of Jacobi, Laguerre, and Hermite and, in general, orthogonal polynomials that satisfy higher-order differential equations of the form

$$\sum_{r=1}^{2m} a_r(x)y^{(r)}(x) = \lambda y(x). \tag{1.10}$$

For example, the classical Jacobi polynomials are orthogonal with respect to the inner products

$$(p,q)_J = \int_{-1}^{1} p(x)\bar{q}(x)(1-x)^{\alpha}(1+x)^{\beta}\,dx \qquad (\alpha > -1, \beta > -1),$$

and

$$(p,q)_{J,W} = (p,q)_J + \int_{-1}^{1} p'(x)\bar{q}'(x)(1-x)^{\alpha+1}(1+x)^{\beta+1}\,dx.$$

Similarly, the Laguerre-type polynomials, which satisfy a fourth-order equation of the form (1.10), are orthogonal with respect to the two inner products

$$(p,q)_{\sigma} = \int_{[0,\infty)} p(x)\bar{q}(x)\,d\sigma(x)$$

and

$$(p,q)_{\sigma,W} = (p,q)_{\sigma} + \int_{0}^{\infty} \{x^2e^{-x}p''(x)\bar{q}''(x) + ((2A+2)x+2)e^{-x}p'(x)\bar{q}'(x)\}\,dx,$$

where A is a fixed, positive constant and σ is the Borel measure generated from the monotonic increasing function $\hat{\sigma} : [0,\infty) \to \mathbb{R}$ defined by

$$\hat{\sigma}(x) = \begin{cases} -1/A & \text{if } x = 0, \\ 1 - e^{-x} & \text{if } x > 0. \end{cases}$$

For additional examples and references, see [5] and [6]. Because of this peculiarity, we shall henceforth assume that the polynomials orthogonal with respect to the inner product (1.5) are *not* orthogonal with respect to the leading measure μ_0.

In §2 of this paper, we shall state and prove a measure theoretic result which is essential in the proof of Theorem 1. Section 3 contains a proof of Theorem 1 while §4 considers several examples to illustrate the theory. Lastly, in §5, we discuss refinements of the theorem with applications to semiclassical orthogonal polynomials.

2. A preliminary measure theory result. The following theorem is essential in establishing the proof of Theorem 1.

THEOREM 2. *Suppose σ is a finite, positive Borel measure on the Borel subsets $\mathcal{B}(\mathbb{R})$ of the real line $\mathbb{R}$ with* $\operatorname{supp}(\sigma) \subset J$, *where J is some interval on the real line. In addition, suppose $K : \mathbb{R} \to \mathbb{R}$ is a polynomial of degree $m \geq 1$ such that*

(i) $K(x) \geq 0, x \in J$, *and*

(ii) $\int_J K(x)\,d\sigma(x) = 0$.

Then there exists an integer $P \geq 1$, *points* $x_1, \ldots, x_P \in J$ *satisfying* $K(x_j) = 0, j = 1, \ldots, P$, *and positive constants* $\alpha_1, \ldots, \alpha_P$ *such that*

$$\operatorname{supp}(\sigma) = \{x_1, \ldots, x_P\}$$

and

$$\sigma = \sum_{j=1}^{P} \alpha_j \delta_{x_j}. \tag{2.1}$$

Proof. If $K(x)$ has no roots in J, we can write $J = \cup_{n=1}^{\infty} A_n$, where

$$A_n := \{x \in J | K(x) > 1/n\} \in \mathcal{B}(\mathbb{R}) \qquad (n = 1, 2, 3, \ldots).$$

However, $0 = \int_J K\,d\sigma \geq \int_{A_n} K\,d\sigma \geq (1/n)\sigma(A_n)$ so that $\sigma(A_n) = 0, n = 1, 2, \ldots$, which forces $\sigma(J) = 0$, contradicting the fact that σ is a positive measure with $\operatorname{supp}(\sigma) \subset J$. Consequently, let $\emptyset \neq R := \{y_1, \ldots, y_r\}$ denote the set of distinct roots of $K(x)$ in J. By repeating the above argument, we can show that $\sigma(B) = 0$ whenever $B \in \mathcal{B}(\mathbb{R})$ and $B \cap R = \emptyset$. If $B \in \mathcal{B}(\mathbb{R})$ is arbitrary, we may write $B = (B \setminus R_B) \cup R_B$, where $R_B := R \cap B$. Then $\sigma(B \setminus R_B) = 0$, and thus

$$\sigma(B) = \sigma(R_B) = \sum_{y_j \in R_B} \alpha_j,$$

where $\alpha_j := \sigma(\{y_j\}), j = 1, \ldots, r$. In particular, $\sigma(J) = \sum_{j=1}^{r} \alpha_j$, and since $\sigma(J) > 0$, at least one of the $\alpha_j's$ is positive. Let $S := \{y_j \in R | \alpha_j > 0\}$ so $S = \operatorname{supp}(\sigma)$ and $m \geq r \geq \operatorname{card}(S) := P \geq 1$. For convenience, rewrite $S = \{x_1, \ldots, x_P\}$ where $\sigma(\{x_j\}) = \alpha_j, j = 1, \ldots, P$. Then

$$\sigma(B) = \sum_{x_j \in B} \alpha_j. \tag{2.2}$$

The proof is complete on noting that (2.2) is equivalent to (2.1). □

3. The Proof of Theorem 1. We start with the following basic result.

LEMMA 3. *Suppose there exists a polynomial* $h : \mathbb{R} \to \mathbb{R}$ *of degree* $m \geq 1$ *such that*

$$(hp, q)_W = (p, hq)_W \qquad (p, q \in \mathcal{P}). \tag{3.1}$$

Let $\{\phi_n(x)\}_{n=0}^{\infty}$ *be the orthogonal polynomial sequence generated by* $(\cdot, \cdot)_W$. *Then* $\{\phi_n(x)\}$ *satisfies the* $(2m+1)$*-term recurrence relation*

$$h(x)\phi_n(x) = \sum_{k=n-m}^{n+m} b_{n,k}\phi_k(x) \qquad (n = m, m+1, \ldots), \tag{3.2}$$

where

$$b_{n,k} = \frac{(h\phi_n, \phi_k)_W}{(\phi_k, \phi_k)_W} \qquad (n = m, m+1, \ldots; k = n-m, n-m+1, \ldots, n+m). \tag{3.3}$$

Proof. For $n \geq m$, we can find real constants $b_{n,k}, k = 0, 1, \ldots, n+m$ such that

$$h(x)\phi_n(x) = \sum_{k=0}^{n+m} b_{n,k}\phi_k(x).$$

For $0 \leq r \leq n+m$, the orthogonality of $\{\phi_n(x)\}$ yields

$$(h\phi_n, \phi_r)_W = \sum_{k=0}^{n+m} b_{n,k}(\phi_k, \phi_r)_W = b_{n,r}(\phi_r, \phi_r)_W. \tag{3.4}$$

On the other hand, $(h\phi_n, \phi_r)_W = (\phi_n, h\phi_r)_W = 0$ if $m + r < n$; i.e. $b_{n,r} = 0$ if $r = 0, \ldots, n-m-1$.

This establishes (3.2), and (3.3) follows from (3.4). □

Remark 8. If the polynomials $\{\phi_n(x)\}$ satisfy a recurrence relation of the form (3.2), with $b_{n,n+m} \neq 0, n \geq m$, then we shall say that $\{\phi_n(x)\}$ satisfies a recurrence relation of *length* $2m+1$. We remark that, in general, it can be quite difficult to compute the recurrence coefficients $b_{n,k}$ given in (3.3). See §5 for a further discussion of this. □

THEOREM 4. *Consider the inner product* $(\cdot,\cdot)_W$, *defined in* (1.5), *where the measures* $\mu_k, k = 0, 1, \ldots, N$, *are positive Borel measures on the Borel subsets of the real line. If there exists a polynomial* $h : \mathbb{R} \to \mathbb{R}$ *of degree* $m \geq 1$ *such that* (3.1) *holds for all* $p, q \in \mathcal{P}$, *then*

$$d\mu_k(x) = \sum_{j=1}^{P(k)} \alpha_{k,j}\delta(x - x_{k,j})\,dx \qquad (k = 1, \ldots, N), \tag{3.5}$$

for some positive integer $P(k)$, *where* $\alpha_{k,j} \geq 0, k = 1, \ldots, N; j = 1, \ldots, P(k)$,

$$R_k = \{x_{k,j}\}_{j=1}^{P(k)} \text{ are the distinct real roots of } h^{(k)}(x), \tag{3.6}$$

and

$$\operatorname{supp}(\mu_k) \subset \bigcap_{j=1}^{k} R_j \neq \emptyset \qquad (k = 1, 2, \ldots, N). \tag{3.7}$$

Proof. Suppose $(hp, q)_W = (p, hq)_W$ for all polynomials $p, q \in \mathcal{P}$. That is,

$$\sum_{k=1}^{N}\sum_{j=0}^{k-1} \binom{k}{j} \int_{\mathbb{R}} h^{(k-j)}[p^{(j)}\bar{q}^{(k)} - p^{(k)}\bar{q}^{(j)}]\,d\mu_k = 0 \qquad (p, q \in \mathcal{P}). \tag{3.8}$$

Substituting $p(x) = 1$ and $q(x) = h(x)$ into (3.8) yields

$$\int_{\mathbb{R}} (h'(x))^2\,d\mu_1 + \int_{\mathbb{R}} (h''(x))^2\,d\mu_2 + \cdots + \int_{\mathbb{R}} (h^{(N)}(x))^2\,d\mu_N = 0.$$

Since $\mu_1, \ldots, \mu_N$ are all positive measures, we must have

$$\int_{\mathbb{R}} (h^{(k)}(x))^2\,d\mu_k = 0 \qquad (k = 1, \ldots, N). \tag{3.9}$$

By Theorem 2, we see that for each $k, 1 \leq k \leq N$, there exists an integer $P(k) \geq 1$, distinct real numbers $x_{k,1}, \ldots, x_{k,P(k)}$ and nonnegative constants $\alpha_{k,1}, \ldots, \alpha_{k,P(k)}$ such that

(i) $\operatorname{supp}(\mu_k) \subset \{x_{k,1}, \ldots, x_{k,P(k)}\}$;

(ii) $h^{(k)}(x_{k,j}) = 0, j = 1, \ldots, P(k)$; i.e., $\{x_{k,j}\}_{j=1}^{P(k)}$ are the distinct real roots of $h^{(k)}(x)$;

(iii) $d\mu_k(x) = \sum_{j=1}^{P(k)} \alpha_{k,j}\delta(x - x_{k,j})\,dx, k = 1, \ldots, N$.

This establishes (3.5) and (3.6) of the theorem.

We now prove, by induction, that

$$\int_{\mathbb{R}} (h^{(k-i)}(x))^2\,d\mu_k = 0 \qquad (i = 0, \ldots, k-1; k = i+1, \ldots, N). \tag{3.10}$$

As a consequence of this induction, we will see that

$$\operatorname{supp}(\mu_k) \subset \bigcap_{i=1}^{k} R_1 \qquad (k = 1, \ldots, N). \tag{3.11}$$

Notice that we have established (3.10) in the case of $i = 0$ (see (3.9)). Suppose, then, that for some $r, 0 \leq r < N$, we have

$$\int_{\mathbb{R}} (h^{(k-i)}(x))^2\,d\mu_k = 0 \qquad (i = 0, \ldots, r; k = i+1, \ldots, N). \tag{3.12}$$

We must show that

$$\int_{\mathbb{R}} (h^{(k-r-1)}(x))^2\,d\mu_k = 0 \qquad (k = r+2, \ldots, N). \tag{3.13}$$

For each $k \in \{1, \ldots, N\}$ and $i \in \{0, \ldots, k-1\}$, we may write

$$h^{(k-i)}(x) = Q_{k-i}(x)(x - x_{k-i,1})^{r_{k-i,1}}(x - x_{k-i,2})^{r_{k-i,2}} \cdots (x - x_{k-i,P(k-i)})^{r_{k-i,P(k-i)}}, \tag{3.14}$$

where $Q_{k-i}(x)$ has no real roots and $r_{k-i,j}, j = 1, \ldots, P(k-i)$, are positive integers. From (3.5), we see that (3.12) implies that

$$\sum_{j=1}^{P(k)} \alpha_{kj} Q_{k-i}^2(x_{kj})(x_{kj} - x_{k-i,1})^{2r_{k-i,1}} \cdots (x_{kj} - x_{k-i,P(k-i)})^{2r_{k-i,P(k-i)}} = 0 \tag{3.15}$$
$$(i = 0, \ldots, r; k = i+1, \ldots, N).$$

Since $Q_{k-i}(x_{k,j}) \neq 0$, either $\alpha_{k,j} = 0$ or $x_{k,j}$ is a root of $h^{(k-i)}(x), i = 0, \ldots, r; k = i+1, \ldots, N, j = 1, \ldots, P(k)$. In other words,

$$\operatorname{supp}(\mu_k) \subset R_{k-i} \qquad (i = 0, 1, \ldots, r; k = i+1, \ldots, N),$$

from which it follows that

$$\operatorname{supp}(\mu_k) \subset \bigcap_{i=1}^{k} R_i \qquad (k = 1, \ldots, r+1) \tag{3.16}$$

and

$$\text{supp}\,(\mu_k) \subset \bigcap_{i=k-r}^{k} R_i \qquad (k = r+2, \ldots, N). \tag{3.17}$$

Moreover, (3.5) and (3.12) also imply that

$$\int_{\mathbb{R}} p(x) h^{(k-i)}(x)\, d\mu_k = 0 \qquad (p \in \mathcal{P}; i = 0, \ldots, r; k = i+1, \ldots, N),$$

and consequently (3.8) simplifies to

$$\sum_{k=r+2}^{N} \sum_{j=r+1}^{k-1} \binom{k}{j} \int_{\mathbb{R}} h^{(k-j)} [p^{(j)} \bar{q}^{(k)} - p^{(k)} \bar{q}^{(j)}]\, d\mu_k = 0 \qquad (p, q \in \mathcal{P}). \tag{3.18}$$

Substitute $p(x) = x^{r+1}/(r+1)!$ and $q(x) = x^{n+r+2}/(n+1)_{r+2}$ into (3.18), where $(a)_n$ is the Pochhammer symbol, to get

$$\begin{aligned} &\binom{r+2}{r+1} \int_{\mathbb{R}} h'(x) x^n\, d\mu_{r+2} + \binom{r+3}{r+1} \int_{\mathbb{R}} h''(x) n x^{n-1}\, d\mu_{r+3} + \cdots \\ &\quad + \binom{N}{r+1} \int_{\mathbb{R}} h^{(N-r-1)}(x) n(n-1) \cdots (n-N+r+3) x^{n-N+r+2}\, d\mu_N = 0. \end{aligned}$$

If we write $h'(x) = \sum_{n=0}^{m-1} \beta_n x^n$ and substitute into the above equation, we immediately obtain

$$\int_{\mathbb{R}} (h'(x))^2\, d\mu_{r+2} = \int_{\mathbb{R}} (h''(x))^2\, d\mu_{r+3} = \cdots = \int_{\mathbb{R}} (h^{(N-r-1)}(x))^2\, d\mu_N = 0,$$

thereby establishing (3.13).

Furthermore, using the representations (3.14) and (3.5), we see that (3.13) yields

$$\begin{aligned} &\sum_{j=1}^{P(k)} \alpha_{kj} Q_{k-r-1}^2(x_{kj}) (x_{kj} - x_{k-r-1,1})^{2r_{k-r-1,1}} \\ &\qquad \cdots (x_{kj} - x_{k-r-1,P(k-r-1)})^{2r_{k-r-1,P(k-r-1)}} = 0 \\ &\qquad (k = r+2, \ldots, N), \end{aligned} \tag{3.19}$$

from which it follows that

$$\text{supp}\,(\mu_k) \subset R_{k-r-1} \qquad (k = r+2, \ldots, N).$$

In particular,

$$\text{supp}\,(\mu_{r+2}) \subset R_1. \tag{3.20}$$

Combining (3.17) and (3.20), we find that

$$\text{supp}\,(\mu_{r+2}) \subset \bigcap_{i=1}^{r+2} R_i.$$

This completes the induction and establishes (3.7) of the theorem. The proof of the theorem is now complete. □

COROLLARY 5. *Under the assumptions of Theorem 4, the degree m of $h(x)$ must be at least $N+1$.*

Proof. Suppose, to the contrary, that $1 \leq m \leq N$ so that $h^{(m)}(x) \equiv c \neq 0$, for some constant c. From (3.9), however, we find that

$$\int_{\mathbb{R}} (h^{(m)}(x))^2 \, d\mu_m = c^2 \int_{\mathbb{R}} d\mu_m = 0,$$

contradicting the fact that μ_m is a positive measure. □

Remark 9. See Remark 11, below, for precise conditions for when there exists a polynomial $h : \mathbb{R} \to \mathbb{R}$ of degree *exactly* $N+1$ satisfying $(hp, q)_W = (p, hq)_W, p, q \in P$. By Lemma 3, it now follows that the sequence of polynomials that are orthogonal with respect to the inner product $(\cdot, \cdot)_W$ will satisfy a recurrence relation of length *at least* $2N+3$.

Remark 10. Stating Theorem 4 slightly differently, it is interesting to note that if the operator $T : \mathcal{P} \to \mathcal{H}$, defined by

$$T(p(x)) = h(x)p(x), \qquad p \in \mathcal{D}(T) = \mathcal{P},$$

is Hermitian, then the measures $\mu_k, 1 \leq k \leq N$, defined in the inner product $(\cdot, \cdot)_W$ necessarily satisfy conditions (3.5) and (3.7). As we shall see below, in Theorem 6, the converse statement is also true.

Summarizing the results of this section so far, we know if there exists a real-valued polynomial h of degree $m \geq 1$ such that (3.1) holds, then the corresponding orthogonal polynomials generated by $(\cdot, \cdot)_W$ satisfy a $(2m+1)$-term recurrence relation; that is, *the degree m of $h(x)$ determines the length of the recurrence relation.* Of course, if h does satisfy (3.1), it will not be unique. Indeed,

$$(hPp, q)_W = (p, hPq)_W \qquad (p, q \in \mathcal{P}),$$

where $P : \mathbb{R} \to \mathbb{R}$ is an arbitrary polynomial. It is natural, then, to ask: what is a polynomial $H : \mathbb{R} \to \mathbb{R}$ of *minimal* degree for which (3.1) holds? From Lemma 3, this minimal polynomial will be *optimal* in the sense that it will yield the recurrence relation of *minimal length* for the associated orthogonal polynomials.

To answer this question, write

$$\mu_k = \sum_{j=1}^{Q(k)} \beta_{kj} \delta_{y_{k,j}} \qquad (k = 1, \ldots, N), \tag{3.21}$$

where we shall now assume that $\beta_{kj} > 0$; i.e.,

$$\operatorname{supp}(\mu_k) = \{y_{k,1}, \ldots, y_{k,Q(k)}\} \qquad (k = 1, \ldots, N), \tag{3.22}$$

and $\operatorname{card}(\operatorname{supp}(\mu_k)) := Q(k) \leq P(k)$.

We now define sets $B_k \subset \operatorname{supp}(\mu_k)$ with $\alpha(k) := \operatorname{card}(B_k) \leq Q(k)$ and polynomials $h_k, 1 \leq k \leq N$ as follows:

(i) Let $B_N = \operatorname{supp}(\mu_N)$ and $h_N(x) = \prod_{y_{N,j} \in B_N} (x - y_{N,j})^N$;

(ii) For $1 \le k \le N-1$, let

$$B_k = \operatorname{supp}(\mu_k) \setminus \bigcup_{j=k+1}^{N} \operatorname{supp}(\mu_j)$$

and define

$$h_k(x) = \begin{cases} 1 & \text{if } B_k = \emptyset, \\ \prod_{y_{kj} \in B_k} (x - y_{kj})^k & \text{if } B_k \neq \emptyset. \end{cases}$$

Notice that the degree of $h_k(x)$ is given by

$$\deg(h_k) = \begin{cases} 0 & \text{if } B_k = \emptyset, \\ k\alpha(k) & \text{if } B_k \neq \emptyset. \end{cases}$$

Finally, let $H(x)$ be the polynomial of degree

$$m(H) := 1 + \sum_{k=1}^{N} \deg(h_k), \tag{3.23}$$

defined by

$$H(x) = \int_0^x \prod_{k=1}^{N} h_k(t)\, dt. \tag{3.24}$$

THEOREM 6. *The polynomial $H : \mathbb{R} \to \mathbb{R}$ defined in* (3.24) *is the unique polynomial (up to a nonzero real constant multiple) of minimal degree satisfying $H(0) = 0$ and* (3.1)*, with the positive measures $\mu_k, 1 \le k \le N$, satisfying conditions* (3.21) *and* (3.22)*. Moreover, if R_k denotes the set of real roots of $H^{(k)}(x)$, then*

$$\operatorname{supp}(\mu_k) \subset \bigcap_{j=1}^{k} R_j \qquad (1 \le k \le N). \tag{3.25}$$

Proof. As in (3.8), we see that

$$(Hp, q)_W - (p, Hq)_W = \sum_{k=1}^{N} \sum_{j=0}^{k-1} \binom{k}{j} \int_{\mathbb{R}} H^{(k-j)} [p^{(j)} \bar{q}^{(k)} - p^{(k)} \bar{q}^{(j)}]\, d\mu_k \qquad (p, q \in \mathcal{P}) \tag{3.26}$$

We show that expression (3.26) is zero for all $p, q \in \mathcal{P}$.

From (3.21) and (3.24), we may write

$$H'(x) = (x - y_{N,1})^N \cdots (x - y_{N,\alpha(N)})^N Q_1(x),$$

where $Q_1(x) = h_1(x) \cdots h_{N-1}(x)$. Hence, for $1 \le r \le N$,

$$\frac{dH^r(x)}{dx^r} = (x - y_{N,1})^{N-r+1} \cdots (x - y_{N,\alpha(N)})^{N-r+1} Q_r(x)$$

for some polynomial $Q_r(x)$. Since the exponents $N - r + 1 \geq 1$ for $1 \leq r \leq N$, it follows from (3.21) that

$$\sum_{j=0}^{N-1} \binom{N}{j} \int_{\mathbb{R}} H^{(N-j)}[p^{(j)}\bar{q}^{(N)} - p^{(N)}\bar{q}^{(j)}]\, d\mu_N = 0 \qquad (p, q \in \mathcal{P}).$$

Now write

$$H'(x) = (x - y_{N,1})^N \cdots (x - y_{N,\alpha(N)})^N \prod (x - y_{N-1,j})^{N-1} P_1(x), \qquad y_{N-1,j} \in B_{N-1}$$

for some $P_1 \in \mathcal{P}$. Then, for $1 \leq r \leq N - 1$, we see that

(3.27)
$$\frac{dH^r(x)}{dx^r} = P_r(x)(x - y_{N,1})^{N-r+1} \cdots (x - y_{N,\alpha(N)})^{N-r+1} \prod (x - y_{N-1,j})^{N-r}, \qquad y_{N-1,j} \in B_{N-1}$$

for some polynomial $P_r(x)$.

Since $\operatorname{supp}(\mu_{N-1}) \subset B_N \cup B_{N-1}$, it follows from (3.27) that

$$\sum_{j=0}^{N-2} \binom{N-1}{j} \int_{\mathbb{R}} H^{(N-j-1)}[p^{(j)}\bar{q}^{(N-1)} - p^{(N-1)}\bar{q}^{(j)}]\, d\mu_{N-1} = 0 \qquad (p, q \in \mathcal{P}).$$

Likewise, and in a similar fashion, we can show that each of the terms on the right-hand side of (3.26) is zero. Hence $H(x)$, as given in (3.24), does satisfy (3.1).

Suppose $g : \mathbb{R} \to \mathbb{R}$ is a polynomial satisfying (3.1), and $g(0) = 0$. From (3.6) and (3.7) of Theorem 4, we see that $g'(x)$ is given by

$$g'(x) = P(x) \prod_{k=1}^{Q(1)} (x - y_{1k}) \prod_{k=1}^{Q(2)} (x - y_{2k})^2 \cdots \prod_{k=1}^{Q(n)} (x - y_{Nk})^N,$$

where $P(x)$ is some polynomial.

From the definition of $H(x)$, it is clear that

(3.28) $$H'(x) \text{ divides } g'(x),$$

i.e., $\deg(g) \geq \deg(H)$. If $\deg(H') = \deg(g')$, then (3.28) yields $KH'(x) = g'(x)$ for some constant $K \neq 0$. Hence $C + KH(x) = g(x)$, for some constant C. Since $H(0) = g(0) = 0$, we must have $C = 0$, and hence

$$g(x) = KH(x).$$

Lastly, we establish the inclusion of (3.25) by induction on $r = N, N - 1, \ldots, 1$.

Let $x_0 \in \operatorname{supp}(\mu_N) = \{y_{N,1}, \ldots, y_{N,Q(N)}\}$. Since

$$H'(x) = (x - y_{N,1})^N \cdots (x - y_{N,Q(N)})^N h_1(x) \cdots h_{N-1}(x),$$

we see that

$$H'(x_0) = H''(x_0) = \cdots = H^{(N)}(x_0) = 0;$$

i.e., $x_0 \in \cap_{j=1}^{N} R_j$.

Suppose then that

$$\operatorname{supp}(\mu_k) \subset \bigcap_{j=1}^{k} R_j \qquad (k = N, N-1, \ldots, r+1).$$

We must show

$$\operatorname{supp}(\mu_r) \subset \bigcap_{j=1}^{r} R_j.$$

By the construction of H, we see that $B_k \subset \cap_{j=1}^{k} R_j$, and hence from our induction hypothesis, we have

$$\begin{aligned}
\operatorname{supp}(\mu_r) &= B_r \cup (\operatorname{supp}(\mu_r) \setminus B_r) \\
&= B_r \cup \left(\operatorname{supp}(\mu_r) \cap \bigcup_{j=r+1}^{N} \operatorname{supp}(\mu_j) \right) \\
&\subset \bigcap_{j=1}^{r} R_j \cup \left(\bigcup_{j=r+1}^{N} \operatorname{supp}(\mu_j) \right) \\
&\subset \bigcap_{j=1}^{r} R_j \cup \bigcap_{j=1}^{r+1} R_j \cup \bigcap_{j=1}^{r+2} R_j \cup \cdots \cup \bigcap_{j=1}^{N} R_j \\
&\subset \bigcap_{j=1}^{r} R_j.
\end{aligned}$$

This completes the induction and finishes the proof of the theorem. □

Remark 11. It is clear that the degree of the polynomial H defined in (3.24) is minimized when $\operatorname{card}(B_k), 1 \le k \le N$, is minimized, and this occurs precisely when $\operatorname{supp}(\mu_k)$ is a singleton set; i.e., $\operatorname{supp}(\mu_k) = \{a\}, 1 \le k \le N$, for some real number a. In this case we see that $H(x)$ is necessarily a nonzero, real constant multiple of $(x-a)^{N+1}$.

By combining the proofs of Lemma 3, Theorem 4, Corollary 5, and Theorem 6, we arrive at a proof of Theorem 1.

4. Examples. In all of the examples below, the measure μ_0 is an arbitrary, positive Borel measure. For other papers concerning recurrence relations for orthogonal polynomials, we refer the reader to [2, Thm. 3.1].

1. It is possible that the polynomial $h(x)$ satisfying (1.6) has only complex roots. For example, if the inner product is given by

$$(p,q)_W = \int_{\mathbb{R}} p(x)\bar{q}(x)\, d\mu_0(x) + \int_{\mathbb{R}} p'(x)\bar{q}'(x)\delta(x)\, dx,$$

then $(hp,q)_W = (p,hq)_W, p,q \in \mathcal{P}$, where $h(x) = x^2 + 1$. For this example, we note that the minimal polynomial $H(x)$, defined in (3.25), is given by $H(x) = x^2$.

2. From Theorem 1, it follows that there does not exist an inner product $(\cdot,\cdot)_W$ given by (1.5) with $N \ge 1$ such that (3.1) is satisfied for

$$h(x) = x^3 + 3x + 3.$$

Indeed, since $h'(x) = 3x^2 + 3$, we see that $R_1 = \emptyset$, contradicting (c) of Theorem 1. In fact, from (3.9), we find that

$$\int_{\mathbb{R}} (3x^2 + 3)^2 \, d\mu_1(x) = 0,$$

which, of course, implies that $\mu_1 \equiv 0$. It is interesting to note that even if we relax our requirements in Theorem 1 and allow the measure μ_1 to be the zero measure, there are still no inner products of the form (1.5) *with* $N \geq 1$ and *with at least one of the measures* $\mu_k, 1 \leq k \leq N$, *nontrivial* for which (3.1) holds with $h(x) = x^3 + 3x + 3$. To see this, observe that since $\deg(h) = 3$, an argument similar to that given in Theorem 1 implies that *necessarily* $N = 1$ or $N = 2$. Since we ruled out the case $N = 1$ above, the only other possibility is that the inner product has the form

$$(p, q)_W = \int_{\mathbb{R}} p(x)\bar{q}(x) \, d\mu_0 + \int_{\mathbb{R}} p''(x)\bar{q}''(x) \, d\mu_2.$$

Assuming then that $(hp, q)_W = (p, hq)_W$ for all polynomials $p, q \in \mathcal{P}$, we find that

$$\int_{\mathbb{R}} (2h'(p'\bar{q}'' - p''\bar{q}') + h''(p\bar{q}'' - p''\bar{q})) \, d\mu_2 = 0 \qquad (p, q \in \mathcal{P}). \tag{4.1}$$

Substituting

$$p(x) = 1 \quad \text{and} \quad q(x) = \frac{x^{n+2}}{(n+1)(n+2)}$$

into this equation yields

$$\int_{\mathbb{R}} h''(x)x^n \, d\mu_2 = 0 \qquad (n \in \mathbb{N}_0),$$

and, hence, it follows that

$$\int_{\mathbb{R}} (h''(x))^2 \, d\mu_2 = \int_{\mathbb{R}} 36x^2 \, d\mu_2 = 0.$$

Of course, since μ_2 is assumed to be nontrivial, we must have

$$d\mu_2(x) = c\delta(x) \, dx$$

for some real constant $c \neq 0$. However, if we substitute $p(x) = x$ and $q(x) = x^2$ into (4.1) we find that $12c = 0$ and thus arrive at a contradiction. We leave it to the reader to formulate and prove the analogue of Theorem 1 when the measures $\mu_k, 1 \leq k \leq N$, are assumed to be nonnegative with at least one of these measures necessarily being positive.

3. Suppose that the Sobolev inner product is given by

$$(p, q)_W = \int_{\mathbb{R}} p(x)\bar{q}(x) \, d\mu_0(x) + \int_{\mathbb{R}} p'(x)\bar{q}'(x)\delta(x) \, dx + \int_{\mathbb{R}} p''(x)\bar{q}''(x)(\delta(x) + \delta(x-1)) \, dx.$$

We now construct the minimal polynomial $H : \mathbb{R} \to \mathbb{R}$ described in Theorem 6. We find that $B_2 = \{0, 1\}$, $B_1 = \emptyset$, $h_2(x) = x^2(x-1)^2$, and $h_1(x) = 1$. Hence the polynomial $H(x)$ is any nonzero, real multiple of

$$12x^5 - 30x^4 + 20x^3 = 60 \int_0^x h_1(t)h_2(t)\, dt.$$

Notice that, in the notation of Theorem 6, $R_1 = \{0, 1\}$, $R_2 = \{0, \frac{1}{2}, 1\}$ and

$$\{0\} = \operatorname{supp}(\mu_1) \subset R_1, \qquad \{0, 1\} = \operatorname{supp}(\mu_2) \subset R_1 \cap R_2.$$

4. Suppose that $(hp, q)_W = (p, hq)_W$, $(p, q \in \mathcal{P}$, where

$$(p, q)_W = \int_{\mathbb{R}} p(x)\bar{q}(x)\, d\mu_0(x) + \int_{\mathbb{R}} p'(x)\bar{q}'(x)\, d\mu_1(x) + \int_{\mathbb{R}} p''(x)\bar{q}''(x)\, d\mu_2(x)$$

and $h(x) = x^5 - 10x^3 + 20x^2 - 15x - 158$. Now $h'(x) = 5(x-1)^3(x+3)$ and $h''(x) = 20(x-1)^2(x+2)$ so that

$$R_1 = \{-3, 1\} \quad \text{and} \quad R_2 = \{-2, 1\}.$$

By Theorem 1, we see that

$$\operatorname{supp}(\mu_1) \subset R_1 \quad \text{and} \quad \operatorname{supp}(\mu_2) \subset R_1 \cap R_2 = \{1\}.$$

Hence, there exists a constant $k > 0$, and nonnegative constants c_1 and c_2 with $c_1^2 + c_2^2 \neq 0$ such that

$$d\mu_2(x) = k\delta(x-1)\, dx, \qquad d\mu_1(x) = (c_1\delta(x-1) + c_2\delta(x+3))\, dx.$$

Since the degree of $h(x)$ is 5, Lemma 3 says that the sequence of polynomials that are orthogonal with respect to $(\cdot, \cdot)_W$ will satisfy an eleven-term recurrence relation. Actually, the algorithm outlined in Theorem 6 indicates that the polynomials will satisfy either a nine-term or a seven-term recurrence relation. Indeed, if $-3 \in \operatorname{supp}(\mu_1)$, then the algorithm outlined in Theorem 6 dictates that

$$h_2(x) = (x-1)^2, \qquad h_1(x) = (x+3),$$

and hence that the minimal polynomial $H(x)$ is any nonzero, real multiple of

$$\int_0^x (t-1)^2(t+3)\, dt.$$

For example, $H(x) = 3x^4 + 4x^3 - 30x^2 + 36x$ will suffice. Consequently, the associated orthogonal polynomials will satisfy a nine-term recurrence relation. On the other hand, if $-3 \notin \operatorname{supp}(\mu_1)$, then $\operatorname{supp}(\mu_1) = \operatorname{supp}(\mu_2) = \{1\}$, and a calculation reveals that $H(x)$ can be taken to be $(x-1)^3$; in this case the polynomials will satisfy a seven-term recurrence relation.

5. Suppose $N = 3$ and

$$\begin{aligned} d\mu_1(x) &= (c_1\delta(x) + c_2\delta(x-1) + c_3\delta(x-2) + c_4\delta(x-3) + c_5\delta(x-4))\, dx, \\ d\mu_2(x) &= (c_6\delta(x) + c_7\delta(x-1) + c_8\delta(x-2))\, dx, \\ d\mu_3(x) &= (c_9\delta(x) + c_{10}\delta(x-1))\, dx, \end{aligned}$$

where $c_i > 0, i = 1, 2, \ldots, 10$. Proceeding as in the previous example, we find that

$$h_3(x) = x^3(x-1)^3,$$
$$h_2(x) = (x-2)^2,$$
$$h_1(x) = (x-3)(x-4),$$

and hence the polynomial $H(x)$ of minimal degree satisfying (3.1) and $H(0) = 0$ is any multiple of

$$1260x^{11} - 1940x^{10} + 123200x^9 - 419265x^8 + 829620x^7 - 960960x^6 + 609840x^5 - 166320x^4.$$

The corresponding orthogonal polynomials will satisfy a 23-term recurrence relation. We emphasize that this length is minimal!

5. Further modifications and applications of Theorem 1. In this section, we shall suppose that the sequence $\{\phi_n(x)\}_{n=0}^{\infty}$ is a monic orthogonal polynomial sequence (MOPS) with respect to the inner product in (1.5). We shall also find it convenient to decompose the inner product (1.5),

$$(p, q)_W = \sum_{k=0}^{N} (p^{(k)}, q^{(k)})_{\mu_k},$$

where the inner products $(\cdot, \cdot)_{\mu_k}$ are defined by

$$(p, q)_{\mu_k} = \int_{\mathbb{R}} p(x)\bar{q}(x)\, d\mu_k \qquad (k = 0, 1, \ldots, N).$$

Let $\{q_n(x)\}_{n=0}^{\infty}$ be the sequence of monic polynomials that is orthogonal with respect to the leading inner product $(\cdot, \cdot)_{\mu_0}$. Writing

$$\phi_n(x) = q_n(x) + \sum_{j=0}^{n-1} \sigma_{n,j} q_j(x), \tag{5.1}$$

we find, in the usual way, the Fourier coefficients $\sigma_{n,j}, 0 \le j \le n-1$:

$$\sigma_{n,j} = \frac{(\phi_n, q_j)_{\mu_0}}{(q_j, q_j)_{\mu_0}} = \frac{-\sum_{k=1}^{N} (\phi_n^{(k)}, q_j^{(k)})_{\mu_k}}{(q_j, q_j)_{\mu_0}} = -\sum_{k=1}^{N} \left(\sum_{s=1}^{Q(k)} \frac{\beta_{k,s} \phi_n^{(k)}(y_{k,s}) q_j^{(k)}(y_{k,s})}{(q_j, q_j)_{\mu_0}} \right).$$

Then (5.1) becomes

$$\phi_n(x) = q_n(x) - \sum_{k=1}^{N} \left(\sum_{s=1}^{Q(k)} \beta_{k,s} \phi_n^{(k)}(y_{k,s}) K_{n-1}^{(0,k)}(x, y_{k,s}) \right), \tag{5.2}$$

where

$$K_{n-1}^{(j,k)}(x, y) = \sum_{l=0}^{n-1} \frac{q_l^{(j)}(x) q_l^{(k)}(y)}{(q_l, q_l)_{\mu_0}}.$$

From the Christoffel–Darboux formula (see [3, p. 23]),

$$K_{n-1}^{(0,0)}(x,y) = \frac{1}{(q_{n-1},q_{n-1})_{\mu_0}} \frac{q_n(x)q_{n-1}(y) - q_{n-1}(x)q_n(y)}{x-y},$$

and taking derivatives with respect to y at the point $y_{k,s}$, we obtain

$$\begin{aligned} K_{n-1}^{(0,k)}(x,y_{k,s}) = \frac{1}{(q_{n-1},q_{n-1})_{\mu_0}} & \left(q_n(x) \sum_{j=0}^{k} \binom{k}{j} \frac{q_{n-1}^{(j)}(y_{k,s})(k-j)!}{(x-y_{k,s})^{k-j+1}} \right. \\ & \left. - q_{n-1}(x) \sum_{j=0}^{k} \binom{k}{j} \frac{q_n^{(j)}(y_{k,s})(k-j)!}{(x-y_{k,s})^{k-j+1}} \right) \\ = \frac{k!(x-y_{k,s})^{-k-1}}{(q_{n-1},q_{n-1})_{\mu_0}} & (q_n(x)T_k(x;y_{k,s};n-1) - q_{n-1}(x)T_k(x;y_{k,s};n)), \end{aligned} \tag{5.3}$$

where $T_k(x;y_{k,s};n)$ denotes the Taylor polynomial of degree k at $y_{k,s}$ for the polynomial $q_n(x)$.

If we denote

$$\tilde{q}_n^{(s)} = \begin{pmatrix} q_n^{(s)}(y_{s,1}) \\ \vdots \\ q_n^{(s)}(y_{s,Q(s)}) \end{pmatrix}, \quad \tilde{K}_{n-1}^{(0,s)} = \begin{pmatrix} K_{n-1}^{(0,s)}(x,y_{s,1}) \\ \vdots \\ K_{n-1}^{(0,s)}(x,y_{s,Q(s)}) \end{pmatrix}, \quad D = \operatorname{diag}(\Lambda_1,\Lambda_2,\ldots,\Lambda_N),$$

where $\Lambda_j = \operatorname{diag}(\beta_{j,1}^{-1},\beta_{j,2}^{-1},\ldots,\beta_{j,Q(j)}^{-1})$, $j=1,2,\ldots,N$, and $\mathcal{R}_{n-1}$ is the block matrix whose diagonal elements are $K_{n-1}^{(i,i)}(y_{i,r},y_{i,s}))_{r,s=1}^{Q(i)}$ while the (l,m) block element is

$$(K_{n-1}^{(l,m)}(y_{l,r},y_{m,s}))_{\substack{r=1,\ldots,Q(l),\\ s=1,\ldots,Q(m)}},$$

then following the same techniques used in [1] and [12], the following result can be deduced.

PROPOSITION 7. *For each* $n \in \mathbb{N}_0$,

(a) *the matrix* $(D+\mathcal{R}_n)$ *is nonsingular*;

(b)
$$\frac{(\phi_n,\phi_n)_W}{(q_n,q_n)_{\mu_0}} = \frac{\det(D+\mathcal{R}_n)}{\det(D+\mathcal{R}_{n-1})};$$

(c)
$$\phi_n(x) = \frac{\begin{vmatrix} q_n(x) & \vdots & (\tilde{K}_{n-1}^{(0,1)})^T \cdots (\tilde{K}_{n-1}^{(0,N)})^T \\ \tilde{q}_n^{(1)} & \vdots & \\ \vdots & \vdots & D+\mathcal{R}_{n-1} \\ \tilde{q}_n^{(N)} & \vdots & \end{vmatrix}}{\det(D+\mathcal{R}_{n-1})},$$

where $(\tilde{K}_{n-1}^{(0,j)})^T$ *denotes the transpose of the vector* $\tilde{K}_{n-1}^{(0,j)}$.

Using (5.3), the following result is an immediate consequence of the above proposition; see also [1, eq. (2.12)] and [12, eq. (2.11)].

COROLLARY 8. *There exists a polynomial $g(x)$ such that*

$$g(x)\phi_n(x) = A_{N,n}(x)q_n(x) + B_{N,n}(x)q_{n-1}(x). \tag{5.4}$$

Indeed,

$$g(x) = \prod_{k=1}^{N} g_k(x), \quad \textit{where}$$

$$g_k(x) = \begin{cases} 1 & \textit{if } B_k = \emptyset, \\ \displaystyle\prod_{y_{kj}\in B_k} (x - y_{kj})^{k+1} & \textit{if } B_k \neq \emptyset, \end{cases}$$

and $A_{N,n}, B_{N,n}$ are polynomials depending on n and, with degrees, respectively, $\deg(g)$ *and* $\deg(g) - 1$.

We remark that, in general, $\deg(g) \geq \deg(h)$. From (5.4), we can explicitly find the parameters $b_{n,k}$ which appear in the recurrence relation (3.2) with some suitable normalization. For example, if $\{\phi_n\}$ is a sequence of monic orthogonal polynomials, multiplication by $g(x)$ in (3.2) yields

$$h(x)g(x)\phi_n(x) = \sum_{k=n-m}^{n+m} b_{n,k}g(x)\phi_k(x) \tag{5.5}$$

and

$$h(x)\left[\sum_{j=n-p}^{n+p} c_{n,j}q_j(x)\right] = \sum_{k=n-m}^{n+m} b_{n,k}\left[\sum_{l=k-p}^{k+p} c_{k,l}q_l(x)\right],$$

where the coefficients $\{c_{n,j}\}$ are the Fourier coefficients of $g(x)\phi_n(x)$ with respect to the orthogonal system $\{q_n\}$. But, from the three-term recurrence relation which is satisfied by the MOPS $\{q_n\}$ and by iteration, we have

$$h(x)q_j(x) = \sum_{l=j-m}^{j+m} d_{j,l}q_l(x),$$

and finally

$$\sum_{j=n-p}^{n+p}\sum_{l=j-m}^{j+m} c_{n,j}d_{j,l}q_l(x) = \sum_{k=n-m}^{n+m}\sum_{l=k-p}^{k+p} b_{n,k}c_{k,l}q_l(x).$$

For fixed n, notice that $2m + 1$ parameters $b_{n,k}$ appear in (5.5). We adopt the convention that

$$\begin{aligned} b_{n,k} &= 0 \quad \text{if } |n-k| > m, \\ c_{k,l} &= 0 \quad \text{if } |k-l| > p, \\ d_{j,l} &= 0 \quad \text{if } |j-l| > m. \end{aligned}$$

Then

$$\sum_{j=n-p-m}^{n+p+m} c_{n,j}d_{j,l} = \sum_{k=n-m-p}^{n+p+m} b_{n,k}c_{k,l} \tag{5.6}$$

for $n-m-p \leq l \leq n+m+p$.

In particular, because of $b_{n,n+m}=1$, we can give the matrix representation

$$C_{2m,p}^{(n)} \begin{pmatrix} b_{n,n+m-1} \\ \vdots \\ b_{n,n-m} \end{pmatrix} = D_{2p,m}^{(n)} \begin{pmatrix} c_{n,n+p} \\ \vdots \\ c_{n,n-p} \end{pmatrix} \tag{5.7}$$

where

$$C_{2m,p}^{(n)} = \begin{pmatrix} c_{n+m-1,n+m-1-p} \cdots & c_{n-m,n+m-1-p} \\ \vdots & \vdots \\ c_{n+m-1,n-m-p} & c_{n-m,n-m-p} \end{pmatrix},$$

$$D_{2p,m}^{(n)} = \begin{pmatrix} d_{n+p,n+m-1-p} \cdots & d_{n-p,n+m-1-p} \\ \vdots & \vdots \\ d_{n+p,n-m-p} & d_{n-p,n-m-p} \end{pmatrix}.$$

However, since $c_{ij}=0$ for $|i-j|>p$ and $c_{i,i-p}>0, i=n+m-1,\ldots,n-m$, we have that $C_{2m,p}^{(n)}$ is a nonsingular upper triangular matrix. As a consequence, we can obtain an explicit representation of the parameters $\{b_{n,k}\}$ of the recurrence relation for which the sequence $\{\phi_n\}$ satisfies in terms of the known parameters $\{c_{i,j}\}$ and $\{d_{i,j}\}$

$$\begin{pmatrix} b_{n,n+m-1} \\ \vdots \\ b_{n,n-m} \end{pmatrix} = (C_{2m,p}^{(n)})^{-1} D_{2p,m}^{(n)} \begin{pmatrix} c_{n,n+p} \\ \vdots \\ c_{n,n-p} \end{pmatrix}.$$

An interesting application of the work in this section is related to the situation when some differential properties of the MOPS $\{q_n\}$ are known. If μ_0 is a semiclassical measure, i.e., if the linear functional u defined on $\mathcal{P}$ by

$$\langle u, p \rangle = \int_{\mathbb{R}} p(x)\, d\mu_0$$

satisfies a distributional equation

$$D(A(x)u) = B(x)u,$$

where A and B are polynomials with $\deg(B) \geq 1$, and D is the derivative operator, it is known (see [13]) that

$$A(x)q_n'(x) = \sum_{j=n-t-1}^{n+r-1} \gamma_{n,j} q_j(x), \tag{5.8}$$

where $\deg(A) := r, \deg(B) := k$, and $t := \max\{r-2, k-1\}$. This last expression, called a structural relation, leads to a second-order linear differential equation

$$A(x,n)q_n''(x) + B(x,n)q_n'(x) + C(x,n)q_n(x) = 0, \tag{5.9}$$

which the MOPS $\{q_n\}$ satisfies. We remark that (5.8) is a weaker condition than (5.9) but, if $\{q_n\}$ satisfies a three-term recurrence relation, then (5.8) is equivalent to (5.9)

(see [7]). However, as the following result shows, there exist nonstandard sequences of MOPS which do not satisfy a three-term recurrence relation such that (5.9) holds.

PROPOSITION 9. *If* $\{q_n\}$ *is a semiclassical* MOPS, *then* $\{\phi_n\}$ *satisfies*

$$\tilde{A}(x,n)\phi_n''(x) + \tilde{B}(x,n)\phi_n'(x) + \tilde{C}(x,n)\phi_n(x) = 0,$$

where $\tilde{A}, \tilde{B}$, *and* $\tilde{C}$ *are polynomials of degrees independent of* n.

Proof. From (5.4), we find

$$(g(x)\phi_n(x))' = A'_{N,n}(x)q_n(x) + B'_{N,n}(x)q_{n-1}(x) + A_{N,n}(x)q_n'(x) + B_{N,n}(x)q_{n-1}'(x).$$

Multiplying by $A(x)$ in the above expression, taking into account the structural relation (5.8) and the three-term recurrence relation for the MOPS $\{q_n(x)\}$, we obtain

$$\begin{aligned} A(x)(g(x)\phi_n(x))' &= S(x,n)q_n(x) + R(x,n)q_{n-1}(x) \\ &= \tilde{S}(x,n)q_n(x) + \tilde{R}(x,n)q_n'(x), \end{aligned} \tag{5.10}$$

where S and R are polynomials and $\tilde{S}$ and $\tilde{R}$ are rational functions.

From (5.4) and (5.10), we obtain

$$\begin{aligned} q_n(x) &= \frac{\begin{vmatrix} g(x)\phi_n(x) & B_{N,n}(x) \\ A(x)(g(x)\phi_n(x))' & R(x,n) \end{vmatrix}}{\begin{vmatrix} A_{N,n}(x) & B_{N,n}(x) \\ S(x,n) & R(x,n) \end{vmatrix}} \\ &= C(x,n)\phi_n(x) + D(x,n)\phi_n'(x), \end{aligned} \tag{5.11}$$

$$\begin{aligned} q_{n-1}(x) &= \frac{\begin{vmatrix} A_{N,n}(x) & g(x)\phi_n(x) \\ S(x,n) & A(x)(g(x)\phi_n(x))' \end{vmatrix}}{\begin{vmatrix} A_{N,n}(x) & B_{N,n}(x) \\ S(x,n) & R(x,n) \end{vmatrix}} \\ &= E(x,n)\phi_n(x) + F(x,n)\phi_n'(x), \end{aligned} \tag{5.12}$$

where C, D, E, and F are rational functions. Using (5.8) and the recurrence relation for $\{q_n\}$, the result follows from (5.11) and (5.12).

Remark 12. The reader will notice that the above proposition includes a constructive approach to constructing the second-order linear differential equation satisfied by the MOPS $\{\phi_n\}$.

Acknowledgments. The second author thanks the first author and the Science and Engineering Research Council (SERC) of Great Britain for the opportunity to visit The University of Wales during the 1991–1992 academic year.

REFERENCES

[1] M. ALFARO, F. MARCELLAN, M. L. REZOLA, AND A. RONVEAUX, *On orthogonal polynomials of Sobolev type: Algebraic properties and zeros*, SIAM J. Math. Anal., 23(1992), pp. 737–757.

[2] H. BAVINCK AND H. G. MEIJER, *On orthogonal polynomials with respect to an inner product involving derivatives: Zeros and recurrence relations*, Indag. Math., N.S. 1(1990), pp. 7–14.

[3] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach Science Publishers, New York, 1978.

[4] A. DURAN, *A generalization of Favard's Theorem for polynomials satisfying a recurrence relation*, J. Approx. Theory, 74(1993), pp. 83–109.

[5] W. N. EVERITT AND L. L. LITTLEJOHN, *Differential operators and the Legendre type polynomials*, Differential and Integral Equations, 1(1988), pp. 97–116.

[6] W. N. EVERITT, A. M. KRALL, L. L. LITTLEJOHN, AND V. P. ONYANGO-OTIENO, *Differential operators and the Laguerre type polynomials*, SIAM J. Math. Anal., 23(1992), pp. 722–736.

[7] W. HAHN, *Über Differentialgleichungen für Orthogonalpolynome*, Monatsh. Math., 95(1983), pp. 269–274.

[8] A. ISERLES, P. E. KOCH, S. P. NØRSETT, AND J. M. SANZ-SERNA, *On polynomials orthogonal with respect to certain Sobolev inner products*, J. Approx. Theory, 65(1991), pp. 151–175.

[9] R. KOEKOEK, *Generalizations of the Classical Laguerre Polynomials and Some q-analogues*, Ph.D. thesis, Technical Univ. of Delft, the Netherlands, 1990.

[10] A. M. KRALL, *Orthogonal polynomials satisfying fourth order differential equations*, Proc. Roy. Soc. Edinburgh, Sec. A, 87(1981), pp. 271–288.

[11] A. KUFNER, *Weighted Sobolev Spaces*, John Wiley, New York, 1985.

[12] F. MARCELLAN AND A. RONVEAUX, *On a class of polynomials orthogonal with respect to a discrete Sobolev inner product*, Indag. Math. New Series 1(1990), pp. 451–464.

[13] P. MARONI, *Une théorie algebrique des polynomes orthogonaux. Application aux polynomes orthogonaux semi-classiques*, in Orthogonal Polynomials and Their Applications, C. Brezinski, L. Gori, and A. Ronveaux, eds., IMACS Annals on Computing and Applied Mathematics, Vol. 9, J. C. Baltzer AG, Basel, Switzerland, 1991, pp. 95–130.

[14] H. G. MEIJER, *Zero distribution of orthogonal polynomials in a certain discrete Sobolev space*, J. Math. Anal. Appl., 172(1993), pp. 520–532.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 468–487, March 1995

012

ON TWO-DIMENSIONAL DEFINITE ORTHOGONAL SYSTEMS AND A LOWER BOUND FOR THE NUMBER OF NODES OF ASSOCIATED CUBATURE FORMULAE*

H. BERENS†, H. J. SCHMID†, AND Y. XU‡

Abstract. In a comprehensive investigation in the 1960s Krall and Sheffer [*Ann. Mat. Pura Appl.*, 76 (1967), pp. 325–376] characterized all bivariate orthogonal polynomial systems which are generated by a second-order differential equation. Actually, they prove that these nine systems are weakly orthogonal and (positive) definite except possibly for two systems. Their paper is completed by showing that these systems are also definite and by determining all parameters for which the classical positive definite systems remain definite. The authors further derive an explicit form of the three-term recursion formulae for all systems. In addition, it is shown that for the associated cubature problem Möller's lower bound applies.

Key words. bivariate orthogonal polynomial systems, Gaussian cubature

AMS subject classifications. 41A10, 41A63, 65D32

1. Introduction.

Let us denote by $\mathbf{P} = \mathbf{R}[x, y]$ the ring of polynomials of two variables with real coefficients, and let $\mathbf{P}_k$, $k = 0, 1, \ldots$, be the linear subspace of $\mathbf{P}$ spanned by

$$1, x, y, \ldots, x^k, \quad x^{k-1}y, \ldots, xy^{k-1}, \quad y^k.$$

Following Krall and Sheffer [5], a monomial basis

$$\{P_j^k\}_{j=0,\ k\in\mathbf{N}}^k, \quad \text{where} \quad P_j^k = x^{k-j}y^j + \text{lower-order terms},$$

is said to be a *weak orthogonal system*, if there exist coefficient matrices

$$C_k = (c_{ij}^k)_{i,j=0,1,\ldots,k}, \qquad \bar{C}_k = (\bar{c}_{ij}^k)_{i=1,2,\ldots,k+1,j=0,1,\ldots,k} \in \mathbf{R}^{k+1\times k+1}$$

and

$$D_k = (d_{i,j}^k)_{i=0,1,\ldots,k,j=0,1,\ldots,k-1}, \qquad \bar{D}_k = (\bar{d}_{ij}^k)_{i=1,2,\ldots,k+1,j=0,1,\ldots,k-1} \in \mathbf{R}^{k+1\times k},$$

such that

$$(1) \quad xP_k = L_{k+1}P_{k+1} + C_kP_k + D_kP_{k-1}, \qquad yP_k = F_{k+1}P_{k+1} + \bar{C}_kP_k + \bar{D}_kP_{k-1},$$

where $P_k = (P_0^k, P_1^k, \ldots, P_k^k)^t$, $k = 0, 1, \ldots$ The matrices L_{k+1} and F_{k+1} are defined as the shift matrices $[E_k\ 0]$ and $[0\ E_k]$, where E_k is the identity in $\mathbf{R}^{k+1\times k+1}$ and $P_{-1} = 0$.

The system is said to be *orthogonal* with respect to the linear functional $\mathcal{I} : \mathbf{P} \to \mathbf{R}$ if, for each $k \in \mathbf{N}_0$, $\mathcal{I}(P_kP_l^t) = 0$, $0 \le l < k$, and if rank $\mathcal{I}(P_kP_k^t) = k + 1$. Here, $P_kP_l^t$ is the tensor product of the vectors P_k and P_l and $\mathcal{I}(P_kP_l^t)$ is the matrix, the coefficients of which are determined by the functional acting on the polynomial coefficients of the tensor product. The matrix $\mathcal{I}(P_kP_k^t)$, $k \in \mathbf{N}_0$, is known as the kth *moment matrix* and is denoted by M_k. Instead of saying $\{P_k\}_{k\in\mathbf{N}_0}$ is an orthogonal

* Received by the editors April 30, 1992; accepted for publication (in revised form) September 7, 1993.

† Mathematical Institute, University of Erlangen–Nuremberg, D-91054 Erlangen, Germany.

‡ Department of Mathematics, University of Oregon, Eugene, Oregon 97403-1222.

system, we also say the system is *definite*; in case the matrices M_k, $k \in \mathbf{N}_0$, are positive definite, we speak of a *positive definite system.*

It is almost obvious that a definite system $\{P_k\}_{k\in\mathbf{N}_0}$ is a weak orthogonal system, i.e., it satisfies the recurrence relations (1). It follows from [11, Thm. 2] that, conversely, *a weak orthogonal system is an orthogonal system with respect to* $\mathcal{I}$, *defined by*

$$\mathcal{I}(P_0^0) = 1 \text{ and } \mathcal{I}(P_j^k) = 0, \qquad 0 \le j \le k,\ k = 1, 2, \ldots,$$

exactly when

$$\text{rank } S_k = k + 1, \quad \text{where } S_k = [D_k\ \bar{D}_k] \in \mathbf{R}^{k+1\times 2k}.$$

Let $\{P_k\}_{k=0}^{\infty}$ be a definite orthogonal system with respect to $\mathcal{I}$. By multiplying (1) by P_{k-1}^t, P_k^t, and P_{k+1}^t, respectively, and applying $\mathcal{I}$, we obtain

$$\text{(2)} \qquad \begin{cases} C_k M_k = \mathcal{I}(xP_kP_k^t), & D_kM_{k-1} = \mathcal{I}(xP_kP_{k-1}^t) = M_kL_k^t, \\ \bar{C}_k M_k = \mathcal{I}(yP_kP_k^t), & \bar{D}_kM_{k-1} = \mathcal{I}(yP_kP_{k-1}^t) = M_kF_k^t. \end{cases}$$

Using these identities, we can compute the moment matrices by induction. Indeed, let $G_k = \text{diag } [2, E_{k-2}]$ and $\bar{G}_k = \text{diag } [E_{k-2}, 2]$; then

$$2\,E_k = L_k^t G_k L_k + F_k^t \bar{G}_k F_k,$$

and consequently,

$$\text{(3)} \qquad 2\,M_k = M_kL_k^tG_kL_k + M_kF_k^t\bar{G}_kF_k = D_kM_{k-1}G_kL_k + \bar{D}_kM_{k-1}\bar{G}_kF_k.$$

Setting $M_0 = 1$, the last equation allows us to compute M_k from M_{k-1}, $k \in \mathbf{N}$.

Again, let $\{P_k\}_{k=0}^{\infty}$ be a definite orthogonal system with respect to $\mathcal{I}$. A cubature formula of degree m for $\mathcal{I}$ is a linear combination of point evaluations

$$K : \mathbf{P} \to \mathbf{R} : p \mapsto K(p) = \sum_{i=0}^{N} \kappa_i\ p(x_i, y_i), \quad \kappa_i \neq 0,\ (x_i, y_i) \in \mathbf{R}^2$$

such that $K(p) = \mathcal{I}(p)$ for all $p \in \mathbf{P}_m$, and at least one $p_0 \in \mathbf{P}_{m+1}$ satisfies $K(p_0) \neq \mathcal{I}(p_0)$. Of special interest are formulae with a minimal number N of nodes. Such formulae are said to be *interpolatory*; i.e., the nodes have the interpolation property. A lower bound for N is

$$N \ge \dim \mathbf{P}_{[m/2]+1} = ([m/2] + 1)([m/2] + 2)/2.$$

The proof is usually given for positive definite systems; the proof in Stroud's book [9, Thm. 3.15-1], however, depends only on the regularity of the moment matrices.

In 1976, Möller [6] improved the bound for strictly positive linear functionals, which nevertheless remains true in the definite case since only orthogonality is involved.

If N is the number of nodes in a cubature formula of degree $m = 2k - 1$ for $\mathcal{I}$, then

$$N \ge \dim \mathbf{P}_{k-1} + \text{rank } M_{k-1}^{\star}/2,$$

where the coefficients of $M^\star_{k-1}$ *are given by*

$$m^\star_{ij} = \mathcal{I}(P^k_{i-1}P^k_j - P^k_i P^k_{j-1}), \qquad i,j = 1,2,\ldots,k.$$

Since $M^\star_{k-1}$ is skew-symmetric, its rank is even. The computation of the coefficients of $M^\star_{k-1}$ and of its rank is difficult in general, and particularly difficult if the integral is not a tensor product of one-dimensional integrals. In the *centrally symmetric case*, where $\mathcal{I}(x^{k-i}y^i) = 0$, $i = 0,1,\ldots,k$, for all odd $k \in \mathbf{N}$, Möller proved

$$\text{rank } M^\star_{k-1} = \begin{cases} k, & \text{if } k \text{ is even,} \\ k-1, & \text{if } k \text{ is odd,} \end{cases} \tag{4}$$

without making use of the explicit form of the matrix.

For the Lebesgue integral over the triangle $\{(x,y) \in \mathbf{R}^2 : 0 \le x,y,1-x-y\}$ Möller [6] computed the rank of $M^\star_{k-1}$ up to $k = 6$ and verified the rank condition (4) for this case. Rasputin [7] then proved it for all $k \in \mathbf{N}$. In a separate paper [2], Berens and Schmid further extended the result to all classical Jacobi weight functions. In their approach, the matrix $M^\star_{k-1}$ was rewritten as

$$\begin{aligned} M^\star_{k-1} &= L_k M_k F_k^t - F_k M_k L_k^t \\ &= L_k \bar{D}_k M_{k-1} - F_k D_k M_{k-1} = (L_k \bar{D}_k - F_k D_k) M_{k-1} \end{aligned}$$

by again applying the identities (2). Introducing

$$M^{\star\star}_{k-1} = L_k \bar{D}_k - F_k D_k,$$

we clearly have

$$\text{rank } M^\star_{k-1} = \text{rank } M^{\star\star}_{k-1}.$$

In the present paper we study a class of definite bivariate orthogonal polynomial systems which are generated by the following second-order differential equation:

$$\mathcal{L}\omega = -\lambda_k \omega, \qquad \lambda_k \in \mathbf{R},\ k \in \mathbf{N}, \tag{5}$$

where

$$\begin{aligned} \mathcal{L}\omega = {} & (ax^2 + d_1 x + e_1 y + f_1)\omega_{xx} + (2axy + d_2 x + e_2 y + f_2)\omega_{xy} \\ & + (ay^2 + d_3 x + e_3 y + f_3)\omega_{yy} + (gx + h_1)\omega_x + (gy + h_2)\omega_y \end{aligned}$$

for some real constants $a \neq 0, g, d_i, e_i, f_i, h_i$, and for

$$\lambda_k = -k((k-1)a + g), \qquad g + ka \neq 0, \quad k = 0,1,\ldots.$$

In their paper, Krall and Sheffer [5] determine all weak orthogonal systems which are generated from (5). Furthermore, they prove that these *nine* systems are indeed either positive definite or definite, except possibly for the systems generated by [5, eqs. (5.22) and (5.53)]. We complete their paper by showing that these systems are also definite, and by determining all parameters for which the classical positive definite systems remain definite. In addition, we shall derive an explicit form of the recursion formula (1) for all these systems, and we will show

$$\text{rank } M^{\star\star}_{k-1} = \begin{cases} k, & \text{if } k \text{ is even,} \\ k-1, & \text{if } k \text{ is odd.} \end{cases} \tag{6}$$

This shows that the lower bound Möller derived for centrally symmetric integrals holds true for all classes considered here. Concerning the existence of cubature formulae of degree $2k-1$ attaining the lower bound, characterizations are known if $\mathcal{I}$ is positive definite and for special functionals even methods are known to construct such formulae; see [8]. In the definite case, however, this is all open. For the existence of cubature formulae of degree $2k-2$, which is not the subject of this paper, we refer to [8] and [11].

Krall and Sheffer were aware of the associated moment problem, i.e., assigning a measure to the functional $\mathcal{I}$ defined by a definite system, in particular a positive measure in case the system is positive definite (Favard's theorem). In the multidimensional case this is quite involved; see Fuglede [3] and the recent results of Xu [10].

2. General approach. We study the nine differential equations separately. To do so, we first determine the coefficients $a_{i\nu}^k$ and $b_{i\nu}^k$ of

$$P_i^k = t_i^k + \sum_{\nu=0}^{k-1} a_{i\nu}^k t_\nu^{k-1} + \sum_{\nu=0}^{k-2} b_{i\nu}^k t_\nu^{k-2} + l.o.t., \qquad t_i^k = x^{k-i}y^i,\ i = 0,1,\dots,k.$$

By applying $\mathcal{L}$ to P_i^k and comparing the coefficients in

$$-\lambda_k \left(t_i^k + \sum_{\nu=0}^{k-1} a_{i\nu}^k t_\nu^{k-1} + \sum_{\nu=0}^{k-2} b_{i\nu}^k t_\nu^{k-2} \right) = \mathcal{L}(t_i^k) + \sum_{\nu=0}^{k-1} a_{i\nu}^k \mathcal{L}(t_\nu^{k-1}) + \sum_{\nu=0}^{k-2} b_{i\nu}^k \mathcal{L}(t_\nu^{k-2}) + l.o.t.,$$

we can compute $a_{i\nu}^k$ and $b_{i\nu}^k$, respectively. Elements $a_{i,j}^k$ and $b_{i,j}^k$, which are not defined explicitly, are equal to zero. In all cases the orthogonal polynomials are of the form

$$P_i^k = t_i^k + \sum_{\nu=-1}^{1} a_{i,i+\nu}^k t_{i+\nu}^{k-1} + \sum_{\nu=-2}^{2} b_{i,i+\nu}^k t_{i+\nu}^{k-2} + l.o.t.,$$

which leads to the recursion formulae

$$(7)\quad \begin{cases} P_i^{k+1} = xP_i^k - \sum_{\nu=-1}^{1} c_{i,i+\nu} P_{i+\nu}^k, - \sum_{\nu=-2}^{2} d_{i,i+\nu} P_{i+\nu}^{k-1}, & i = 0,1,\dots,k, \\ P_i^{k+1} = yP_{i-1}^k - \sum_{\nu=-1}^{1} \bar{c}_{i,i+\nu} P_{i+\nu}^k - \sum_{\nu=-2}^{2} \bar{d}_{i,i+\nu} P_{i+\nu}^{k-1}, & i = 1,2,\dots,k+1. \end{cases}$$

To simplify the notation, we omitted and henceforth will omit the superindex k in $c_{i,j}$ and $d_{i,j}$, respectively. Again, by comparing the coefficients in

$$P_i^{k+1} - xP_i^k = \sum_{\nu=-1}^{1} (a_{i,i+\nu}^{k+1} - a_{i,i+\nu}^k) t_{i+\nu}^k + \sum_{\nu=-2}^{2} (b_{i,i+\nu}^{k+1} - b_{i,i+\nu}^k) t_{i+\nu}^{k-1} + l.o.t.$$

$$= -\sum_{\nu=-1}^{1} c_{i,i+\nu} P_{i+\nu}^k + \sum_{\nu=-1}^{1} c_{i,i+\nu} \sum_{\rho=-1}^{1} a_{i+\nu,i+\nu+\rho}^k P_{i+\nu+\rho}^{k-1} + \sum_{\nu=-2}^{2} (b_{i,i+\nu}^{k+1} - b_{i,i+\nu}^k) P_{i+\nu}^{k-1},$$

we obtain

$$\begin{cases} c_{i,i+\nu} = a_{i,i+\nu}^k - a_{i,i+\nu}^{k+1}, & \nu = -1,0,1, \\ d_{i,i+\nu} = b_{i,i+\nu}^k - b_{i,i+\nu}^{k+1} - \sum_{\rho=-1}^{1} c_{i,i+\rho} a_{i+\rho,i+\nu}^k, & \nu = -2,-1,0,1,2. \end{cases}$$

Analogously, we find

$$\begin{cases} \bar{c}_{i,i+\nu} = a^k_{i-1,i+\nu-1} - a^{k+1}_{i,i+\nu}, & \nu = -1, 0, 1, \\ \bar{d}_{i,i+\nu} = b^k_{i-1,i+\nu-1} - b^{k+1}_{i,i+\nu} - \sum_{\rho=-1}^{1} \bar{c}_{i,i+\rho} a^k_{i+\rho,i+\nu}, & \nu = -2, -1, 0, 1, 2. \end{cases}$$

Hence, we can compute the (possibly) nonvanishing elements of the real $k+1 \times k$ matrices

$$D_k = \begin{bmatrix} d_{0,0} & d_{0,1} & d_{0,2} & & & \\ d_{1,0} & d_{1,1} & d_{1,2} & \ddots & & \\ d_{2,0} & d_{2,1} & d_{2,2} & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & d_{k-3,k-2} & d_{k-3,k-1} \\ & & \ddots & \ddots & d_{k-2,k-2} & d_{k-2,k-1} \\ & & & \ddots & d_{k-1,k-2} & d_{k-1,k-1} \\ & & & & d_{k,k-2} & d_{k,k-1} \end{bmatrix}$$

and

$$\bar{D}_k = \begin{bmatrix} \bar{d}_{1,0} & \bar{d}_{1,1} & \bar{d}_{1,2} & \bar{d}_{1,3} & & & \\ \bar{d}_{2,0} & \ddots & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & \ddots & \bar{d}_{k-3,k-1} \\ & & & \ddots & \ddots & \ddots & \bar{d}_{k-2,k-1} \\ & & & & \ddots & \ddots & \bar{d}_{k-1,k-1} \\ & & & & & \ddots & \bar{d}_{k,k-1} \\ & & & & & & \bar{d}_{k+1,k-1} \end{bmatrix}.$$

As stated above, a definite system will be obtained if and only if

$$\operatorname{rank} S_k = \operatorname{rank} [D_k \ \bar{D}_k] = k + 1.$$

In order to prove (6), we have to determine the elements of $M^{\star\star}_{k-1}$:

$$m^{\star\star}_{ij} = \bar{d}_{i,j-1} - d_{i,j-1}, \qquad i, j = 1, 2, \ldots, k.$$

All rank conditions can be attacked directly if the matrices D_k, $\bar{D}_k$ in the recursion formula are explicitly known. To be on the safe side we checked the results by using a Maple program for all nine systems and transferred the results directly into the text. The input is a function defining $a^k_{i,j}$ and $b^k_{i,j}$ for the system in question.

3. Explicit recursion formulae and the rank conditions. In this section we present the nine differential equations, numbered as in [5], the corresponding image polynomials of the monomials under $\mathcal{L}$, the highest coefficients of the corresponding orthogonal polynomials of degree k, and the coefficients of the recursion formulae. Finally, in all cases the conditions on the free parameters are discussed in order to satisfy the rank condition for S_k as well as $M^{\star\star}_{k-1}$. Since we are only interested in

determining the rank of S_k and of $M^{\star\star}_{k-1}$, respectively, we will delete common factors in the rows and columns without changing the notation.

Equation (3.10). Differential equation:

$$\omega_{xx} + \omega_{yy} - x\omega_x - y\omega_y = -k\omega.$$

Monomials:

$$\mathcal{L}(t_i^k) = -kt_i^k + (k-i)(k-i-1)t_i^{k-2} + i(i-1)t_{i-2}^{k-2}.$$

Nonvanishing coefficients $a^k_{i,j}$ and $b^k_{i,j}$:

$$b^k_{i,i-2} = -i(i-1)/2,\ b^k_{i,i} = -(k-i)(k-i-1)/2.$$

Nonvanishing coefficients in the recursion:

$$d_{i,i} = k-i, \qquad \bar{d}_{i,i-2} = i-1.$$

Nonvanishing elements of $M^{\star\star}_{k-1}$:

$$m^{\star\star}_{i,i-1} = i-1, \quad i = 2,3,\dots,k, \quad m^{\star\star}_{i,i+1} = -(k-i), \quad i = 1,2,\dots,k-1.$$

The differential equation defines the product *Hermite* polynomial system (the tensor product of the Hermite polynomials); it is definite, even positive definite, and the rank condition (6) holds.

Equation (3.12). Differential equation:

$$x\omega_{xx} + y\omega_{yy} + (1+\alpha-x)\omega_x + (1+\beta-y)\omega_y = -k\omega.$$

Monomials:

$$\mathcal{L}(t_i^k) = -kt_i^k + (k-i)(k-i-\alpha)t_i^{k-1} + i(i+\beta)t_{i-1}^{k-1}.$$

Nonvanishing coefficients $a^k_{i,j}$ and $b^k_{i,j}$:

$$\begin{aligned}
a^k_{i,i-1} &= -i(i+\beta),\\
a^k_{i,i} &= -(k-i)(k-i+\alpha),\\
b^k_{i,i-2} &= -i(i-1)(i+\beta)(i+\beta-1)/2,\\
b^k_{i,i-1} &= i(i+\beta)(k-i)(k-i+\alpha),\\
b^k_{i,i} &= -(k-i)(k-i-1)(k-i-\alpha)(k-i-\alpha-1)/2.
\end{aligned}$$

Nonvanishing coefficients in the recursion:

$$c_{i,i} = 2k+1-2i+\alpha,\ \bar{c}_{i,i-1} = 2i-1+\beta,\ d_{i,i} = (k-i)(k-i+\alpha),\ \bar{d}_{i,i-2} = (i-1)(i-1+\beta).$$

The matrix S_k:

$$\begin{bmatrix}
k(k+\alpha) & & & 0 & & \\
& (k-1)(k+\alpha+1) & & \beta+1 & & \\
& & \ddots & & 2(\beta+2) & \\
& & & \alpha+1 & & \ddots \\
& & & 0 & & k(k+\beta)
\end{bmatrix};$$

it has rank $k+1$ exactly when $-\alpha, -\beta \notin \mathbf{N}$.

Nonvanishing elements of $M^{\star\star}_{k-1}$:

$$m^{\star\star}_{i,i-1} = (i-1)(i-1+\beta),\ i = 2,3,\ldots,k,\ m^{\star\star}_{i,i+1} = -(k-i)(k-i+\alpha),\ i = 1,2,\ldots,k-1.$$

The differential equation defines the product *Laguerre* polynomial system (the tensor product of the Laguerre polynomials); it is definite if and only if $-\alpha, -\beta \notin \mathbf{N}$, in which case the rank condition (6) also holds. For $\alpha, \beta > -1$ the system is even positive definite.

Equation (3.13). Differential equation:

$$\omega_{xx} + y\omega_{yy} - x\omega_x + (1+\alpha-y)\omega_y = -k\omega.$$

Monomials:

$$\mathcal{L}(t^k_i) = -kt^k_i + i(i+\alpha)t^{k-1}_{i-1} + (k-i)(k-i-1)t^{k-2}_i.$$

Nonvanishing coefficients $a^k_{i,j}$ and $b^k_{i,j}$:

$$a^k_{i,i-1} = -i(i+\alpha),\ b^k_{i,i-2} = (i+\alpha)(i+\alpha-1)i(i-1)/2,\ b^k_{i,i} = -(k-i)(k-i-1)/2.$$

Nonvanishing coefficients in the recursion:

$$\bar{c}^k_{i,i-1} = 2i+\alpha-1, \quad d_{i,i} = k-i, \quad \bar{d}_{i,i-2} = (i-1)(i-1+\alpha).$$

The matrix S_k:

$$\begin{bmatrix} k & & & & 0 & & & & \\ & k-1 & & & \alpha+1 & & & & \\ & & \ddots & & & 2(\alpha+2) & & & \\ & & & \ddots & & & \ddots & & \\ & & & & 1 & & & \ddots & \\ & & & & 0 & & & & k(k+\alpha) \end{bmatrix};$$

rank $S_k = k+1$ exactly when $-\alpha \notin \mathbf{N}$. The classical systems satisfy $\alpha > -1$.

Nonvanishing elements of $M^{\star\star}_{k-1}$:

$$m^{\star\star}_{i,i-1} = (i-1)(i-1+\alpha), \quad i = 2,3,\ldots,k, \quad m^{\star\star}_{i,i+1} = -(k-i),\ i = 1,2,\ldots,k-1.$$

The differential equation defines the product *Hermite–Laguerre* polynomial system (the tensor product of Hermite polynomials and Laguerre polynomials); it is definite if and only if $-\alpha \notin \mathbf{N}$, and in this case the rank condition (6) also holds.

Equation (5.14). Differential equation:

$$(x^2-1)\omega_{xx} + 2xy\omega_{xy} + (y^2-1)\omega_{yy} + gx\omega_x + gy\omega_y = k(k+g-1)\omega,\ -g \notin \mathbf{N}_0.$$

Monomials:

$$\mathcal{L}(t^k_i) = k(k+g-1)t^k_i - (k-i)(k-i-1)t^{k-2}_i - i(i-1)t^{k-2}_{i-2}.$$

Nonvanishing coefficients $a^k_{i,j}$ and $b^k_{i,j}$:

$$b\ b^k_{i,i-2} = -i(i-1), \qquad b\ b^k_{i,i} = -(k-i)(k-i-1),$$

where $b = 2(2k + g - 3)$.

Nonvanishing coefficients in the recursion:

$$d\ d_{i,i-2} = i(i-1), \qquad d\ d_{i,i} = (k-i)(k+g+i-2),$$
$$d\ \bar{d}_{i,i-2} = (i-1)(2k+g-i-1), \quad d\ \bar{d}_{i,i} = (k-i+1)(k-i),$$

where $d = (2k + g - 3)(2k - 1 + g)$. Since $d_{i,i-2} \neq 0$ and

$$\det \begin{bmatrix} d_{00} & 0 & \bar{d}_{11} \\ 0 & \bar{d}_{20} & 0 \\ d_{20} & 0 & \bar{d}_{31} \end{bmatrix} \neq 0,$$

we obtain rank $S_k = k + 1$.

Nonvanishing elements of $M^{\star\star}_{k-1}$:

$$m^{\star\star}_{i,i-1} = \frac{i-1}{2k+g-3}, \quad i = 2, 3, \ldots, k, \quad m^{\star\star}_{i,i+1} = -\frac{k-i}{2k+g-3}, \quad i = 1, 2, \ldots, k-1.$$

The differential equation defines the *circle-polynomial* system (cf. Chap. VI in [1]); it is definite, even positive definite, and the rank condition (6) holds.

Equation (5.19). Differential equation:

$$3y\omega_{xx} + 2\omega_{xy} - x\omega_x - y\omega_y = -k\omega.$$

Monomials:

$$\mathcal{L}(t^k_i) = -kt^k_i + 3(k-i)(k-i-1)t^{k-1}_{i+1} + 2(k-i)it^{k-2}_{i-1}.$$

Nonvanishing coefficients $a^k_{i,j}$ and $b^k_{i,j}$:

$$a^k_{i,i+1} = -3(k-i)(k-i-1),$$
$$b^k_{i,i-1} = -i(k-i), \qquad b^k_{i,i+2} = 9(k-i-2)(k-i-3)(k-i)(k-i-1)/2.$$

Nonvanishing coefficients in the recursion:

$$c_{i,i+1} = 6(k-i), \quad d_{i,i-1} = i, \quad \bar{d}_{i,i-1} = k+1-i.$$

The rank of S_k is easily determined to be equal to $k + 1$.

As shown in [5], and easily verified by the induction formula (3), the moment matrix M_k, $k \in \mathbf{N}_0$, has cross-diagonal nonvanishing elements:

$$m^{(k)}_{i,k-i} = i!(k-i)!, \qquad 0 \leq i \leq k.$$

Nonvanishing elements of $M^{\star\star}_{k-1}$:

$$m^{\star\star}_{i,i} = k - 2i + 1, \qquad i = 1, 2, \ldots, k.$$

The system is definite, not positive definite, and the rank condition (6) holds.

Equation (5.22). Differential equation:

$$(x^2+y+1)\omega_{xx}+(2xy+2x)\omega_{xy}+(y^2+2y+1)\omega_{yy}+gx\omega_x+gy\omega_y = -k(k-1+g)\omega, \ -g \notin \mathbf{N}_0.$$

Monomials:

$$\mathcal{L}(t_i^k) = k(k+g-1)t_i^k + (k-i)(k-i-1)t_{i+1}^{k-1} + 2i(k-1)t_{i-1}^{k-1} + (k-i)(k-i-1)t_i^{k-2} + i(i-1)t_{i-2}^{k-2}.$$

Nonvanishing coefficients $a_{i,j}^k$ and $b_{i,j}^k$:

$$a_{i,i-1}^k = \frac{2i(k-1)}{2k+g-2},$$
$$a_{i,i+1}^k = \frac{(k-i)(k-i-1)}{2k+g-2},$$
$$b_{i,i-2}^k = \frac{i(i-1)(g+2(k-1)(2k-3))}{2(2k+g-2)(2k+g-3)},$$
$$b_{i,i}^k = \frac{(k-i)(k-i-1)(g+2(i+1)(2k-3))}{2(2k+g-2)(2k+g-3)},$$
$$b_{i,i+1}^k = \frac{(k-i)(k-i-1)(k-i-2)(k-i-3)}{2(2k+g-2)(2k+g-3)}.$$

Nonvanishing coefficients in the recursion:

$$c\, c_{i,i-1} = -2ig,$$
$$c\, c_{i,i+1} = -2(k-i)(k+g-1+i),$$
$$c\, \bar{c}_{i,i-1} = -2ig - 4k^2 - 2gk + 4k + 2g,$$
$$c\, \bar{c}_{i,i+1} = 2(k-i+1)(k-i),$$
$$d\, d_{i,i-2} = ig^2(i-1),$$
$$d\, d_{i,i} = g(k-i)(-g^2+5g+2ig-4gk+2ik-5-3i+9k-4k^2+2i^2),$$
$$d\, d_{i,i+2} = (2g+3k-3+i)(-k+i+2)(-k+i+1)(-k+i),$$
$$d\, \bar{d}_{i,i-2} = g^2(-g-2k+1+i)(i-1),$$
$$d\, \bar{d}_{i,i} = g(2k+g-2i-2)(k-i+1)(k-i),$$
$$d\, \bar{d}_{i,i+2} = (k-i-2)(k-i+1)(k-i-1)(k-i),$$

where $c = (2k+g-2)(2k+g)$ and $d = (2k-1+g)(2k+g-3)(2k+g-2)^2$. The entries $d_{i,i-2}$, $i = 2, 3, \ldots, k$, do not vanish. To determine rank S_k it sufficies to determine the rank of the submatrix

$$A = \begin{bmatrix} d_{0,0} & 0 & \bar{d}_{1,1} \\ 0 & \bar{d}_{2,0} & 0 \\ d_{2,0} & 0 & \bar{d}_{3,1} \end{bmatrix}.$$

Here, $\det A = -4kg^5(2k+g-2)(2k+g-3)^2(2k+g-4)$. We obtain a definite orthogonal system if $g \notin \mathbf{N}_0$; this was left open in [5].

It follows from the representation of D_k and $\bar{D}_k$ and formula (3) that the coefficient $m_{k,k}^{(k)}$ of the moment matrix M_k is given by

$$m_{k,k}^{(k)} = \bar{d}_{k+1,k-1}^{(k)}\, m_{k-1,k-1}^{(k-1)}, \qquad k \in \mathbf{N},$$

where

$$\bar{d}_{k+1,k-1}^{(k)} = \frac{-kg^2(g+k-2)}{(g+2k-3)(g+2k-2)^2(g+2k-1)}.$$

Since $M_0 = 1$, $m_{k,k}^{(k)} \neq 0$ for a given admissible g and for all $k \in \mathbf{N}_0$. Assume the system to be positive definite; then $m_{k,k}^{(k)} > 0$ for all $k \in \mathbf{N}_0$. An inductive argument implies that $g < -(2k-1)$ for all $k \in \mathbf{N}$, which is obviously impossible.

Nonvanishing elements of $M_{k-1}^{\star\star}$:

$$
\begin{aligned}
m\, m_{i,i-1}^{\star\star} &= -g^2(i-1), \qquad i = 2,3,\ldots,k,\\
m\, m_{i,i+1}^{\star\star} &= -g(-g+3+3i-3k)(k-i), \qquad i = 1,2,\ldots,k-2,\\
m\, m_{i,i+3}^{\star\star} &= 2(k-i-2)(k-i-1)(k-i), \qquad i = 1,2,\ldots,k-4,
\end{aligned}
$$

where $m = (2k+g-3)(2k+g-2)^2$.

Hence, the system is definite and not positive definite for $g \notin \mathbf{N}_0$, and equation (6) holds.

Equation (5.52). Differential equation:

$$
\begin{aligned}
&(x^2-x)\omega_{xx} + 2xy\omega_{xy} + (y^2-y)\omega_{yy}\\
&+((\alpha+1)x - \beta - 1)\omega_x + ((\alpha+1)y - \gamma - 1)\omega_y = k(k+\alpha)\omega, \quad -\alpha \notin \mathbf{N}_0.
\end{aligned}
$$

Monomials:

$$
\mathcal{L}(t_i^k) = k(k+\alpha)t_i^k - (k-i)(k-i+\beta)t_i^{k-1} - i(i+\gamma)t_{i-1}^{k-1}.
$$

Nonvanishing coefficients $a_{i,j}^k$ and $b_{i,j}^k$:

$$
\begin{aligned}
a_{i,i}^k &= \frac{-(k-i)(k-i+\beta)}{2k+\alpha-1},\\
a_{i,i-1}^k &= \frac{-i(i+\gamma)}{2k+\alpha-1},\\
b_{i,i-2}^k &= \frac{i(i-1)(i+\gamma)(i+\gamma-1)}{2(2k+\alpha-1)(2k+\alpha-2)},\\
b_{i,i-1}^k &= \frac{i(i+\gamma)(k-i)(k-i+\beta)}{(2k+\alpha-1)(2k+\alpha-2)},\\
b_{i,i}^k &= \frac{(k-i-1)(k-i)(k-i+\beta-1)(k-i+\beta)}{2(2k+\alpha-1)(2k+\alpha-2)}.
\end{aligned}
$$

Nonvanishing coefficients in the recursion:

$$
\begin{aligned}
c\, c_{i,i-1} &= -2i(i+\gamma),\\
c\, c_{i,i} &= -1+2i+\alpha-\beta+2\alpha k-2i^2-2i\alpha+2k^2+2i\beta+\beta\alpha,\\
c\, \bar{c}_{i,i-1} &= -2i^2+4ik+2i+2i\alpha-2i\gamma-2k-1-\alpha+2\gamma k+\gamma+\gamma\alpha,\\
c\, \bar{c}_{i,i} &= -2(-k-1+i)(-k-1+i-\beta),\\
d\, d_{i,i-2} &= i(i-1)(i+\gamma)(i-1+\gamma),\\
d\, d_{i,i-1} &= i(-\alpha-2\alpha k+2i\alpha-\beta\alpha+2\beta+2-4i-2i\beta+2k-2k^2+2i^2)(i+\gamma)\\
&= i(i+\gamma)[2(k-i+1)(k-i+\beta+1)-(2k-2i+\beta+1)(2k+\alpha)],\\
d\, d_{i,i} &= (\alpha+k+i-1)(i+\alpha-1+k-\beta)(k-i+\beta)(k-i),\\
d\, \bar{d}_{i,i-2} &= (-\alpha-2k+i)(i-2k-\alpha+\gamma)(i-1)(i-1+\gamma),\\
d\, \bar{d}_{i,i-1} &= [-2i\alpha+\alpha-\gamma\alpha-2\gamma k+2i^2-4ik+2k+2i\gamma](k+1-i)(k+1-i+\beta)\\
&= (k-i+1)(k-i+\beta+1)[2i(i+\gamma)-(2i+\gamma-1)(2k+\alpha)],\\
d\, \bar{d}_{i,i} &= (k+1-i+\beta)(k+1-i)(k-i+\beta)(k-i),
\end{aligned}
$$

where $c = (2k+\alpha-1)(2k+\alpha+1)$ and $d = (2k+\alpha)(2k+\alpha-1)^2(2k+\alpha-2)$.

If $k = 1$ the rank condition is satisfied if and only if

$$\det S_1 = \alpha(\alpha+1)(\beta+1)(\gamma+1)(\alpha-\beta-\gamma-1) \neq 0.$$

Furthermore, in this case rank $M_0^{\star\star} = 0$. Henceforth we shall assume that $k \geq 2$.

In order to obtain rank $S_k = k+1$, the first and last row of S_k must be different from zero. Since $k+\beta$ and $k+\gamma$ are a common factor of the two rows, respectively, we get in addition to $-\alpha \notin \mathbf{N}_0$,

$$-\beta \notin \mathbf{N} \text{ and } -\gamma \notin \mathbf{N}.$$

Thus, $d_{i,i-2}$ and $\bar{d}_{i,i}$ are nonvanishing elements in S_k. The rank condition is satisfied if the determinants of the matrices

$$A = \begin{bmatrix} d_{00} & \bar{d}_{10} & \bar{d}_{11} \\ d_{10} & \bar{d}_{20} & \bar{d}_{21} \\ d_{20} & 0 & \bar{d}_{31} \end{bmatrix} \quad \text{or} \quad B = \begin{bmatrix} d_{k-2,k-2} & 0 & \bar{d}_{k-1,k-1} \\ d_{k-1,k-2} & d_{k-1,k-1} & \bar{d}_{k,k-1} \\ d_{k,k-2} & d_{k,k-1} & \bar{d}_{k+1,k-1} \end{bmatrix}$$

do not vanish. We find

$$\det A = a\ k(k+\beta)(\gamma+1)(\gamma+2)(k+\alpha-\gamma-2)(k+\alpha-\beta-\gamma-2),$$
$$\det B = a\ k(k+\gamma)(\beta+1)(\beta+2)(k+\alpha-\beta-2)(k+\alpha-\beta-\gamma-2),$$

where $a = 2(2k+\alpha-3)(2k+\alpha-2)^2(2k+\alpha-1)$.

Let us first consider the case $\alpha = \beta+\gamma-k+2$. Inserting α into the formulae of $d_{i,j}$ and $\bar{d}_{i,j}$ and deleting the common factor d, we get for D_k,[1] after deleting the common factor $i+\gamma$ in the $(i-1)$st column,

$$\begin{aligned} d_{i+1,i-1} &= i(i+1)(i+\gamma+1), \\ d_{i,i-1} &= i(2(k-i)(k-i+\beta) - (2(k-i)+\beta+1)(k+\beta+\gamma)), \\ d_{i-1,i-1} &= (\beta+\gamma+i)(k-i+\beta+1)(k-i+1), \\ d_{k-1,k-1} &= (\beta+k+\gamma)(\beta+1),\ d_{k,k-1} = -k(\beta+k+\gamma)(\beta+1), \end{aligned}$$

and similarly for $\bar{D}_k$, after deleting the common factor $k-i+\beta$ in the ith column,

$$\begin{aligned} \bar{d}_{i+2,i} &= (k-i+\beta+\gamma)(i+1)(i+\gamma+1), \\ \bar{d}_{i+1,i} &= (k-i)(2i(i+\gamma+1) - (2i+\gamma+1)(k+\beta+\gamma)), \\ \bar{d}_{i,i} &= (k-i+\beta+1)(k-i+1)(k-i), \\ \bar{d}_{1,0} &= -k(k+\beta+\gamma),\ \bar{d}_{2,0} = (k+\beta+\gamma). \end{aligned}$$

Next we add the first $k-1$ columns of D_k to the last $k-1$ columns of $\bar{D}_k$ and obtain for $i = 1, 2, \ldots, k-1$,

$$\begin{aligned} \bar{d}_{i+2,i} &= (i+1)(i+\gamma+1), \\ \bar{d}_{i+1,i} &= -(2i(k-i) + (k-i)(\gamma+1) + i(\beta+1)), \\ \bar{d}_{i,i} &= (k-i+1)(k-i+\beta+1), \\ \bar{d}_{1,0} &= -k, \qquad \bar{d}_{2,0} = 1, \end{aligned}$$

[1] Recall that we do not change the notation when we delete common factors in rows and columns of the matrices under consideration.

where the common factor $k+\beta+\gamma \neq 0$ has been deleted from all elements. Now we subtract the ith column of $\bar{D}_k$ multiplied by i from the $(i-1)$st column of D_k and get

$$\begin{aligned} d_{i+1,i-1} &= 0, \\ d_{i,i-1} &= -i(k-i+\beta+1), \\ d_{i-1,i-1} &= (k-i+1)(k-i+\beta+1), \\ d_{k,k-1} &= -k, \qquad d_{k-1,k-1} = 1; \end{aligned}$$

here, the common factor $(\beta+\gamma+1) \neq 0$ has been deleted. Finally, we subtract for $i = 1, 2, \ldots, k-1$, the $(i-1)$st column of D_k from the ith column of $\bar{D}_k$, which leads to

$$\bar{d}_{i+2,i} = i+1, \quad \bar{d}_{i+1,i} = -(k-i), \quad \bar{d}_{i,i} = 0, \quad \bar{d}_{1,0} = -k, \quad \bar{d}_{2,0} = 1.$$

Thus $\bar{D}_k = D_k$ and rank $S_k = k$. Hence the parameters α, β, γ must in addition satisfy the condition $\beta+\gamma-\alpha \notin \mathbf{N}_0$ in order to get a definite system.

There remains to consider the case when $\beta = \gamma = \alpha + k - 2$. Then

$$\begin{aligned} d_{i+1,i-1} &= i(i+1)(k+i+\alpha-1)(k+i+\alpha-2), \\ d_{i,i-1} &= i(k+i+\alpha-2)(2(k-i+1)(2k-i+\alpha-1) \\ &\qquad -(2k+\alpha)(3k-2i+\alpha-1)), \\ d_{i-1,i-1} &= i(k+i+\alpha-2)(2k-i+\alpha-1)(k-i+1), \end{aligned}$$

where $i(k+i+\alpha-2)$ is a common factor in the $(i-1)$st column of D_k. Similarly,

$$\begin{aligned} \bar{d}_{i+2,i} &= (i+1)(k-i)(k+i+\alpha-1)(2k-i+\alpha-2), \\ \bar{d}_{i+1,i} &= (k-i)(2k-i+\alpha-2)(2(i+1)(k+i+\alpha-1) - (k+2i+\alpha-1)(2k+\alpha)), \\ \bar{d}_{i,i} &= (k-i)(k-i+1)(2k-i+\alpha-2)(2k-i+\alpha-1), \end{aligned}$$

here $(k-i)(2k-i+\alpha-2)$ is a common factor in the ith column of $\bar{D}_k$. Thus, for $i = 1, 2, \ldots, k$,

$$\bar{d}_{i+2,i} - d_{i+1,i-1} = 0, \quad \bar{d}_{i+1,i} - d_{i,i-1} = 0, \quad \bar{d}_{i,i} - d_{i-1,i-1} = 0.$$

By subtracting the first $k-2$ columns in D_k from the last $k-2$ columns of $\bar{D}_k$, we find that the rank of S_k is equivalent to the rank of the following tridiagonal matrix:

$$\begin{bmatrix} a_0 & c_0 & & \\ b_1 & a_1 & \ddots & \\ & \ddots & \ddots & c_{k-1} \\ & & b_k & a_k \end{bmatrix},$$

where

$$\begin{aligned} a_0 &= \bar{d}_{10}, \quad a_i = d_{i,i-1}, \quad i = 1, 2, \ldots, k, \\ b_1 &= \bar{d}_{2,0}, \quad b_i = d_{i,i-2}, \quad i = 2, 3, \ldots, k, \\ c_i &= d_{i,i}, \qquad i = 0, 1, \ldots, k-1. \end{aligned}$$

Since

$$
\begin{aligned}
a_i &= 2(k-i+1)(2k-i+\alpha-1) - (2k+\alpha)(3k-2i+\alpha-1), \qquad i = 0,1,\dots,k,\\
c_i &= (k-i)(2k-i+\alpha-2),\ i = 0,1,\dots,k-1,\\
b_i &= i(k+i+\alpha-2), \qquad i = 1,2,\dots,k,
\end{aligned}
$$

it is not hard to factor the matrix as follows:

$$
\begin{bmatrix}
a_0 & c_0 & & \\
b_1 & a_1 & \ddots & \\
 & \ddots & \ddots & c_{k-1}\\
 & & b_k & a_k
\end{bmatrix}
=
\begin{bmatrix}
\alpha_0 & & & \\
b_1 & \alpha_1 & & \\
 & \ddots & \ddots & \\
 & & b_k & \alpha_k
\end{bmatrix}
\begin{bmatrix}
1 & \gamma_0 & & \\
 & 1 & \ddots & \\
 & & \ddots & \gamma_{k-1}\\
 & & & 1
\end{bmatrix},
$$

where

$$
\begin{aligned}
\alpha_i &= -(2k+\alpha-i-2)(k+\alpha+i-1), \qquad i = 0,1,\dots,k,\\
\gamma_i &= -(k-i)/(k+\alpha+i-1), \qquad i = 0,1,\dots,k-1.
\end{aligned}
$$

Hence, no further restrictions on the parameters α, β, and γ are necessary.

The differential equation defines the *Jacobi polynomial* system on the simplex (cf. Chap. VI in [1]); it is definite if and only if $-\alpha, -\alpha+\beta+\gamma \notin \mathbf{N}_0$ and $-\beta, -\gamma \notin \mathbf{N}$; it is even positive definite for $\alpha > \beta+\gamma+1, \beta > -1, \gamma > -1$.

Nonvanishing elements of M^{**}_{k-1}:

$$
\begin{aligned}
m\, m^{**}_{i,i-1} &= (2k-2i+\alpha-\gamma)(i-1)(i-1+\gamma), \qquad i = 2,3,\dots,k\\
m\, m^{**}_{i,i} &= i(i+\gamma)(2k-2i+\beta+1) - (k-i+1)(k-i+\beta+1)(2i+\gamma-1),\\
&\qquad i = 1,2,\dots,k,\\
m\, m^{**}_{i,i+1} &= -(\alpha-2+2i-\beta)(k-i+\beta)(k-i), \qquad i = 1,2,\dots,k-1,
\end{aligned}
$$

where $m = (2k+\alpha-1)^2(2k+\alpha-2)$. To determine the rank of M^{**}_{k-1}, let us introduce the notation

$$
M^{**} = -\begin{bmatrix}
a_1 & c_1 & & \\
b_2 & a_2 & \ddots & \\
 & \ddots & \ddots & c_{k-1}\\
 & & b_k & a_k
\end{bmatrix},
$$

where

$$
\begin{aligned}
b_i &= -(2k-2i+\alpha-\gamma)(i-1)(i-1+\gamma), \qquad i = 2,3,\dots,k\\
a_i &= (k-i+1)(k-i+\beta+1)(2i+\gamma-1) - i(i+\gamma)(2k-2i+\beta+1),\\
&\qquad i = 1,2,\dots,k,\\
c_i &= (\alpha-2+2i-\beta)(k-i+\beta)(k-i), \qquad i = 1,2,\dots,k-1.
\end{aligned}
$$

Since there is at most one element in the subdiagonal which might vanish, the rank of M^{**} is greater than or equal to $k-2$. Since the rank of M^{**}_{k-1} is even this means rank $M^{**} \geq k-1$ if k is odd.

Denoting the principal minors of M^{**} by A_i, $i = 0, 1, \ldots, k$, we can compute $\det M^{**} = A_k$ recursively via

$$A_0 = 1, \quad A_1 = a_1, \quad A_i = a_i A_{i-1} - b_i c_{i-1} A_{i-2}, \; i = 2, 3, \ldots, k.$$

By introducing the expressions

$$\begin{aligned} S_i &= (k-i)(\gamma+i)(k+\beta-i), \\ T_i &= (2k-2i+\alpha-\gamma-2)(2i+\alpha-\beta-2), \qquad i = 1, 2, \ldots, \end{aligned}$$

we obtain

$$a_i = iS_i - (i-1)S_{i-1} \text{ and } b_i c_{i-1} = -(i-1)S_{i-1}T_{i-1};$$

the recursion for the subdeterminants can then be rewritten as

$$A_0 = 1, \quad A_1 = S_1, \quad A_i = [iS_i - (i-1)S_{i-1}]A_{i-1} + (i-1)S_{i-1}T_{i-1}A_{i-2}, \; i = 2, 3, \ldots, k.$$

In what follows we need the following main identity connecting the S_is and T_is:

$$\begin{aligned} &(m+1)\left[(i-2m+1)S_{i+1} - (i-2m)S_i - S_{2(i-m)+1}\right] \\ &= \frac{(i-2m+1)(i-2m)}{2}\left[T_{i+1} - T_{i-m}\right], \end{aligned}$$

where $m = 0, 1, 2 \ldots$; the identity is easily verified.

By induction we prove that

$$A_i = U_{2i-1} + \sum_{\nu=1}^{[i/2]} \lambda_\nu^i U_{2(i-\nu)-1} T_i T_{i-1} \ldots T_{i-\nu+1}, \qquad i = 1, 2, \ldots, k,$$

where

$$U_{2m-1} = S_1 S_3 \ldots S_{2m-1}, \quad m = 1, 2, \ldots, \quad \text{and } \lambda_\nu^i = \frac{i!}{(i-2\nu)!\nu!2^\nu}.$$

For $i = 1$ we have $a_1 = S_1 = U_1$. By assuming the formula to be true for A_{i-1} and A_i, we obtain for $[(i+1)S_{i+1} - iS_i]A_i + iS_i T_i A_{i-1}$ the expression

$$\begin{aligned} &((i+1)S_{i+1} - iS_i)\left[U_{2i-1} + \sum_{\nu=1}^{[i/2]} \lambda_\nu^i U_{2(i-\nu)-1} T_i T_{i-1} \ldots T_{i-\nu+1}\right] \\ &\qquad + iS_i T_i \left[U_{2i-3} + \sum_{\nu=1}^{[(i-1)/2]} \lambda_\nu^{i-1} U_{2(i-\nu)-3} T_{i-1} T_{i-2} \ldots T_{i-\nu}\right]. \end{aligned}$$

We apply the main identity for $m = 0$ to $[(i+1)S_{i+1} - iS_i]U_{2i-1}$ and get

$$\begin{aligned} &U_{2i+1} + \frac{i(i+1)}{2} U_{2i-1} T_{i+1} - \frac{i(i+1)}{2} U_{2i-3} S_{2i-1} T_i \\ &+ (i+1)S_{i+1} \sum_{\nu=1}^{[i/2]} \lambda_\nu^i U_{2(i-\nu)-1} T_i T_{i-1} \ldots T_{i-\nu+1} + iS_i T_i U_{2i-3} \end{aligned}$$

$$
\begin{aligned}
&+iS_iT_i\sum_{\nu=1}^{[(i-1)/2]}\lambda_\nu^{i-1}U_{2(i-\nu)-3}T_{i-1}T_{i-2}\dots T_{i-\nu}\\
&-iS_iT_i\sum_{\nu=1}^{[i/2]}\lambda_\nu^{i}U_{2(i-\nu)-1}T_{i-1}T_{i-2}\dots T_{i-\nu+1}\\
&=U_{2i+1}+\frac{i(i+1)}{2}U_{2i-1}T_{i+1}\\
&+U_{2i-3}T_i\left[(i+1)\lambda_1^iS_{i+1}-i(\lambda_1^i-1)S_i-\frac{i(i+1)}{2}S_{2i-1}\right]\\
&+(i+1)S_{i+1}\sum_{\nu=2}^{[i/2]}\lambda_\nu^iU_{2(i-\nu)-1}T_iT_{i-1}\dots T_{i-\nu+1}\\
&-iS_iT_i\sum_{\nu=2}^{[i/2]}\lambda_\nu^iU_{2(i-\nu)-1}T_{i-1}T_{i-2}\dots T_{i-\nu+1}\\
&+iS_iT_i\sum_{\nu=2}^{[(i-1)/2]+1}\lambda_{\nu-1}^{i-1}U_{2(i-\nu)-1}T_{i-1}T_{i-2}\dots T_{i-\nu+1}\\
&=U_{2i+1}+\frac{i(i+1)}{2}U_{2i-1}T_{i+1}\\
&+U_{2i-3}T_i\left[(i+1)\lambda_1^iS_{i+1}-i(\lambda_1^i-1)S_i-\frac{i(i+1)}{2}S_{2i-1}\right]\\
&+\sum_{\nu=2}^{[i/2]}U_{2(i-\nu)-1}T_iT_{i-1}\dots T_{i-\nu+1}\left[(i+1)\lambda_\nu^iS_{i+1}-i(\lambda_\nu^i-\lambda_{\nu-1}^{i-1})S_i\right]+R.
\end{aligned}
$$

For i even we have $R=0$, while for i odd, say, $i=2\nu+1$ we obtain

$$R=(2\nu+1)\lambda_\nu^{2\nu}U_{2\nu+1}T_{2\nu+1}\dots T_{\nu+1}.$$

Next we apply the main identity for $m=\nu$ and, by recalling that

$$\lambda_\nu^{i+1}=\frac{(i+1)!}{(i+1-2\nu)!\nu!2^\nu},$$

we obtain

$$
\begin{aligned}
&(i+1)\lambda_\nu^iS_{i+1}-i(\lambda_\nu^i-\lambda_{\nu-1}^{i-1})S_i=\frac{(i+1)!}{(i-2\nu+1)!\nu!2^\nu}\left[(i-2\nu+1)S_{i+1}-(i-2\nu)S_i\right]\\
&=\lambda_\nu^{i+1}\left[S_{2(i-\nu)+1}+\frac{(i-2\nu+1)(i-2\nu)}{2(\nu+1)}(T_{i+1}-T_{i-\nu})\right].
\end{aligned}
$$

Note that the second summand vanishes for i even, say, $i=2\nu$. By setting the undefined term $\lambda_{\nu+1}^{i+1}$ to be equal to 0 for i even and $i=2\nu$, we can rewrite the last equation as

$$(i+1)\lambda_\nu^iS_{i+1}-i(\lambda_\nu^i-\lambda_{\nu-1}^{i-1})S_i=\lambda_\nu^{i+1}S_{2(i-\nu)+1}+\lambda_{\nu+1}^{i+1}(T_{i+1}-T_{i-\nu}).$$

The main identity for $m=1$ can be written as

$$(i-1)S_{i+1}-(i-2)S_i-S_{2i-1}=\frac{(i-1)(i-2)}{4}(T_{i+1}-T_{i-1}),$$

which gives

$$\begin{aligned}&(i+1)\lambda_1^i S_{i+1} - i(\lambda_1^i - 1)S_i - \frac{i(i+1)}{2}S_{2i-1}\\ &= \frac{i(i+1)}{2}[(i-1)S_{i+1} - (i-2)S_i - S_{2i-1}]\\ &= \lambda_2^{i+1}[T_{i+1} - T_{i-1}].\end{aligned}$$

Thus we can simplify the last expression obtained for $[(i+1)S_{i+1} - iS_i]\,A_i + iS_i\,T_iA_{i-1}$ to

$$\begin{aligned}&U_{2i+1} + \lambda_1^{i+1}U_{2i-1}\,T_{i+1} + \lambda_2^{i+1}U_{2i-3}\,T_{i+1}T_i - \lambda_2^{i+1}U_{2i-3}\,T_iT_{i-1}\\ &+ \sum_{\nu=2}^{[i/2]} \lambda_\nu^{i+1}S_{2(i-\nu)+1}U_{2(i-\nu)-1}\,T_iT_{i-1}\ldots T_{i-\nu+1} + \sum_{\nu=2}^{[i/2]} \lambda_{\nu+1}^{i+1}U_{2(i-\nu)-1}\,T_{i+1}T_i\ldots T_{i-\nu+1}\\ &- \sum_{\nu=2}^{[i/2]} \lambda_{\nu+1}^{i+1}U_{2(i-\nu)-1}\,T_iT_{i-1}\ldots T_{i-\nu} + R,\end{aligned}$$

which can be rewritten as

$$\begin{aligned}&U_{2i+1} + \sum_{\nu=0}^{[i/2]} \lambda_{\nu+1}^{i+1}U_{2(i-\nu)-1}\,T_{i+1}T_i\ldots T_{i-\nu+1}\\ &+ \sum_{\nu=2}^{[i/2]} \lambda_\nu^{i+1}U_{2(i-\nu)+1}\,T_iT_{i-1}\ldots T_{i-\nu+1} - \sum_{\nu=1}^{[i/2]} \lambda_{\nu+1}^{i+1}U_{2(i-\nu)-1}\,T_iT_{i-1}\ldots T_{i-\nu} + R;\end{aligned}$$

the last two sums reduce to $-R$, which finally gives

$$U_{2i+1} + \sum_{\nu=1}^{[i/2]+1} \lambda_\nu^{i+1}U_{2(i+1-\nu)-1}\,T_{i+1}T_i\ldots T_{i+1-\nu+1}.$$

Taking into account that for i even, say, $i = 2\nu$ $\lambda_{\nu+1}^{2\nu+1} = 0$ and that for i odd $[i/2] + 1 = [(i+1)/2]$, the last formula completes the induction.

Thus we obtain

$$A_k = S_1S_3\ldots S_{2(k-[k/2])-1} \sum_{\nu=0}^{[k/2]} \lambda_\nu^k S_{2(k-[k/2])+1}S_{k+3}\ldots S_{2(k-\nu)-1}T_kT_{k-1}\ldots T_{k-\nu+1}.$$

Since $S_{2(k-[k/2])-1} = S_k = 0$ for odd k, we get in this case rank $M_{k-1}^{\star\star} = k-1$. To finish the discussion let us assume k to be even. Then $S_1S_3\ldots S_{k-1} \neq 0$. By introducing the shifted factorials (Pochhammer symbols)

$$(a)_0 = 1,\ (a)_n = a(a+1)\ldots(a+n-1), \qquad a \in \mathbf{R},$$

and regarding the identities

$$\begin{aligned}&(a-\nu)_\nu = (-1)^\nu(1-a)_\nu, \qquad (a)_{i-\nu} = (-1)^\nu\frac{(a)_i}{(1-a-i)_\nu},\\ &2^{k/2}\left(\frac{a+1}{2}\right)_{k/2} = (a+1)(a+3)\ldots(a+k-1),\end{aligned}$$

we can fully factorize A_k. Indeed,

$$S_{k+1}S_{k+3}\dots S_{2(k-\nu)-1} = 2^{3(k/2-\nu)}(1/2)_{k/2-\nu}\left(\frac{\gamma+k+1}{2}\right)_{k/2-\nu}\left(\frac{-\beta+1}{2}\right)_{k/2-\nu}$$
$$= 2^{3(k/2-\nu)}(-1)^{\nu}\frac{(1/2)_{k/2}\left(\frac{\gamma+k+1}{2}\right)_{k/2}\left(\frac{-\beta+1}{2}\right)_{k/2}}{\left(\frac{1-k}{2}\right)_{\nu}\left(\frac{-\gamma-2k+1}{2}\right)_{\nu}\left(\frac{\beta-k+1}{2}\right)_{\nu}},$$

similarly

$$T_kT_{k-1}\dots T_{k-\nu+1} = 2^{2\nu}\left(\frac{\alpha-\gamma-2}{2}\right)_{\nu}\left(\frac{\alpha-\beta+2k-2\nu}{2}\right)_{\nu}$$
$$= 2^{2\nu}(-1)^{\nu}\left(\frac{\alpha-\gamma-2}{2}\right)_{\nu}\left(\frac{\beta-\alpha-2k+2}{2}\right)_{\nu},$$

and finally

$$\lambda_{\nu}^{k} = \frac{2^{\nu}}{\nu!}\left(-\frac{k}{2}\right)_{\nu}\left(\frac{1-k}{2}\right)_{\nu}.$$

Thus

$$A_k = S_1S_3\dots S_{k-1}2^{3k/2}\left(\frac{1}{2}\right)_{k/2}\left(\frac{\gamma+k+1}{2}\right)_{k/2}\left(\frac{-\beta+1}{2}\right)_{k/2}\Sigma,$$

where

$$\Sigma = \sum_{\nu=0}^{k/2}\frac{\left(-\frac{k}{2}\right)_{\nu}\left(\frac{\alpha-\gamma-2}{2}\right)_{\nu}\left(\frac{\beta-\alpha-2k+2}{2}\right)_{\nu}}{\nu!\left(\frac{-\gamma-2k+1}{2}\right)_{\nu}\left(\frac{\beta-k+1}{2}\right)_{\nu}}.$$

Since the numerator of Σ vanishes for $\nu > k/2$, we can represent Σ as a generalized hypergeometric series, evaluated at 1; i.e.,

$$\Sigma = {}_3F_2\left(\frac{\alpha-\gamma-2}{2}, \frac{\beta-\alpha-2k+2}{2}, -k/2; \frac{-\gamma-2k+1}{2}, \frac{\beta-k+1}{2}; 1\right). \tag{8}$$

Since the ${}_3F_2$ sum is balanced, we can apply the Pfaff–Saalschütz-formula (see, e.g., [4]) and obtain

$$\Sigma = \frac{\left(\frac{-2k-\alpha+3}{2}\right)_{k/2}\left(\frac{\alpha-\beta-\gamma-1}{2}\right)_{k/2}}{\left(\frac{-\gamma-2k+1}{2}\right)_{k/2}\left(\frac{-\beta+1}{2}\right)_{k/2}} = \frac{\left(\frac{k+\alpha-1}{2}\right)_{k/2}\left(\frac{\alpha-\beta-\gamma-1}{2}\right)_{k/2}}{\left(\frac{\gamma+k+1}{2}\right)_{k/2}\left(\frac{-\beta+1}{2}\right)_{k/2}}.$$

Thus

$$A_k = S_1S_3\dots S_{k-1}2^{3k/2}\left(\frac{1}{2}\right)_{k/2}\left(\frac{k+\alpha-1}{2}\right)_{k/2}\left(\frac{\alpha-\beta-\gamma-1}{2}\right)_{k/2}$$
$$= 2^{3k}\left(\frac{1}{2}\right)_{k/2}^{2}\left(\frac{\gamma+1}{2}\right)_{k/2}\left(\frac{\beta+1}{2}\right)_{k/2}\left(\frac{k+\alpha-1}{2}\right)_{k/2}\left(\frac{\alpha-\beta-\gamma-1}{2}\right)_{k/2}.$$

Hence for the admissible parameters α, β, and γ we get rank $M_{k-1}^{\star\star} = k$ for even k. To summarize, if the system is definite (i.e., $-\alpha, -\alpha+\beta+\gamma \notin \mathbf{N}_0$ and $-\beta, -\gamma \notin \mathbf{N}$), then the rank condition (6) holds. For the positive definite system the proof of the rank condition can be significantly simplified, as done in [2].

Equation (5.53). Differential equation:

$$x^2\omega_{xx} + 2xy\omega_{xy} + (y^2 - y)\omega_{yy} + g(x-1)\omega_x + g(y-\alpha)\omega_y = k(k+g-1)\omega, \ -g \notin \mathbf{N}_0.$$

Monomials:

$$\mathcal{L}(t_i^k) = k(k+g-1)t_i^k - i(i+\alpha g-1)t_{i-1}^{k-1} - g(k-i)t_i^{k-1}.$$

Nonvanishing coefficients $a_{i,j}^k$ and $b_{i,j}^k$:

$$a_{i,i-1}^k = \frac{-i(i-1+\alpha g)}{2k+g-2}, \qquad a_{i,i}^k = \frac{-g(k-i)}{2k+g-2},$$
$$b_{i,i-2}^k = \frac{i(i-1)(i-1+\alpha g)(i-2+\alpha g)}{2(2k+g-2)(2k+g-3)},$$
$$b_{i,i-1}^k = \frac{g(k-i)i(i-1+\alpha g)}{(2k+g-2)(2k+g-3)},$$
$$b_{i,i}^k = \frac{g^2(k-i)(k-i-1)}{2(2k+g-2)(2k+g-3)}.$$

Nonvanishing coefficients in the recursion:

$$c\ c_{i,i-1} = -2i(i-1+\alpha g),$$
$$c\ c_{i,i} = g(g-2+2i),$$
$$c\ \bar{c}_{i,i-1} = 4ik + 2ig - 4k - 2g + 2\alpha gk + \alpha g^2 - 2i^2 + 2i - 2i\alpha g,$$
$$c\ \bar{c}_{i,i} = 2g(-k-1+i),$$
$$d\ d_{i,i-2} = i(i-1)(i-1+\alpha g)(i-2+\alpha g),$$
$$d\ d_{i,i-1} = -ig(g-3+2i)(i-1+\alpha g),$$
$$d\ d_{i,i} = g^2(g-2+i+k)(-k+i),$$
$$d\ \bar{d}_{i,i-2} = (-g+1+i-2k)(i+\alpha g-g-2k)(i-1)(i-2+\alpha g),$$
$$d\ \bar{d}_{i,i-1} = g(k+1-i)(-\alpha g^2 - 2\alpha gk + 2i\alpha g + \alpha g + 2g - 2ig - 2 + 4k - 4ik + 2i^2),$$
$$d\ \bar{d}_{i,i} = g^2(-k-1+i)(-k+i),$$

where $c = (2k+g-2)(2k+g)$ and $d = (2k-1+g)(2k+g-2)^2(2k+g-3)$. To obtain the full rank of S_k, the last row of S_k must be different from zero, i.e., the following elements must not vanish:

$$d_{k,k-2} = k(k-1)(k+\alpha g-1)(k+\alpha g-2),$$
$$d_{k,k-1} = -kg(2k+g-3)(k+\alpha g-1),$$
$$\bar{d}_{k+1,k-1} = (k+g-2)(k+g-\alpha g-1)k(k+\alpha g-1).$$

Hence $-\alpha g \neq 0, 1, 2, \ldots$. Since $\bar{d}_{i,i} \neq 0$, $i = 1, 2, \ldots, k-1$, we consider the submatrix

$$S_k^\star = \begin{bmatrix} \bar{d}_{k-1,k-1} & 0 & d_{k-2,k-2} \\ \bar{d}_{k,k-1} & d_{k-1,k-1} & d_{k-1,k-2} \\ \bar{d}_{k+1,k-1} & d_{k,k-1} & d_{k,k-2} \end{bmatrix},$$

the determinant of which is

$$\det S_k^\star = 2g^4 k(2k+g-2)(2k+g-4)(2k+g-3)^2(k+\alpha g-1).$$

For $-g, -\alpha g \neq 0, 1, 2, \ldots$, a definite system will be obtained independent of α; this was left open by [5].

It follows from the representation of D_k and $\bar{D}_k$ and formula (3) that the coefficient $m_{0,0}^{(k)}$ of the moment matrix M_k is given by

$$m_{0,0}^{(k)} = d_{0,0}^{(k)}\, m_{0,0}^{(k-1)}, \qquad k \in \mathbf{N},$$

where

$$d_{0,0}^{(k)} = \frac{-kg^2(g+k-2)}{(g+2k-3)(g+2k-2)^2(g+2k-1)}.$$

Since $M_0 = 1$, $m_{0,0}^{(k)} \neq 0$ for a given admissible g and for all $k \in \mathbf{N}_0$. Assume that the system is positive definite; then $m_{0,0}^{(k)} > 0$ for all $k \in \mathbf{N}_0$. An inductive argument implies that $g < -(2k-1)$ for all $k \in \mathbf{N}$, which is obviously impossible.

Nonvanishing elements of $M_{k-1}^{\star\star}$:

$$\begin{aligned} m\, m_{i,i-1}^{\star\star} &= -(i-1)(i-2+\alpha g)(2i-3+\alpha g), \qquad i = 2, 3, \ldots, k \\ m\, m_{i,i}^{\star\star} &= g(-2ik+2k-\alpha gk+3i^2-5i+2i\alpha g+2-\alpha g), \\ &\qquad i = 1, 2, \ldots, k, \\ m\, m_{i,i+1}^{\star\star} &= g^2(k-i), \qquad i = 1, 2, \ldots, k-1, \end{aligned}$$

where $m = (2k+g-2)(2k+g-3)$. No element in the superdiagonal vanishes, hence the matrix condition is satisfied. The system is definite and not positive definite if and only if $-g, -\alpha g \notin \mathbf{N}_0$; in this case also the rank condition (6) holds.

Equation (5.55). Differential equation:

$$(x+\alpha)\omega_{xx} + (2y+2)\omega_{xy} + x\omega_x + y\omega_y = k\omega.$$

Monomials:

$$\mathcal{L}(t_i^k) = kt_i^k + (k-i)(k+i-1)t_i^{k-1} + \alpha(k-i)(k-i-1)t_i^{k-2} + 2(k-i)it_{i-1}^{k-2}.$$

Nonvanishing coefficients $a_{i,j}^k$ and $b_{i,j}^k$:

$$a_{i,i}^k = (k-i)(k+i-1),\ b_{i,i-1}^k = (k-i)i,\ b_{i,i}^k = (k-i)(k-i-1)(\alpha+(k+i-1)(k+i-2))/2.$$

Nonvanishing coefficients in the recursion:

$$\begin{aligned} c_{i,i}^k &= -2k, \qquad \bar{c}_{i,i}^k = -(2k-2i+2), \\ d_{i,i-1} &= -i, \qquad d_{i,i} = (k-i)(k+i-1-\alpha), \\ \bar{d}_{i,i-1} &= -(k-i+1), \qquad \bar{d}_{i,i} = (k-i)(k-i+1). \end{aligned}$$

Thus we obtain

$$\begin{bmatrix} k(k-\alpha-1) & & & & & -k & k(k-1) & & & \\ -1 & \ddots & & & & & -(k-1) & \ddots & & \\ & -2 & \ddots & & & & & \ddots & \ddots & \\ & & \ddots & \ddots & & & & & \ddots & 2 \\ & & & \ddots & (2k-\alpha-2) & & & & & -1 \\ & & & & -k & & & & & 0 \end{bmatrix},$$

which is of rank $k+1$ for all α.

In [5] it is shown, and it can be easily verified by the induction formula (3), that the moment matrix M_k, $k \in \mathbf{N}_0$, has vanishing elements $m_{i,j}^{(k)}$ for $k < i+j$ and that the cross-diagonal elements are given by

$$m_{i,k-i}^{(k)} = (-1)^k i!(k-i)!, \qquad 0 \le i \le k.$$

Nonvanishing elements of $M_{k-1}^{\star\star}$:

$$m_{i,i}^{\star\star} = -k + 2i - 1, \; m_{i,i+1}^{\star\star} = -(k-i)(2i-2-\alpha).$$

Hence, the system is definite, not positive definite, for all α; the rank condition (6) is also satisfied.

Acknowledgment. The authors are indebted to Dr. C. Markett for acquainting us with the generalized hypergeometric function ${}_3F_2$ and for pointing out the representation (8) of the sum Σ, which led to the final factorization of A_k. We also thank the referee for his kind and thorough review.

REFERENCES

[1] P. APPELL AND J. K. DE FÉRIET, *Fonctions hypergéometriques et hypersphériques–Polynomes d'Hermite*, Gauthiers–Villars et Cie., Paris, 1926.

[2] H. BERENS AND H. J. SCHMID, *On the number of nodes of odd degree cubature formulae for integrals with Jacobi weights on a simplex*, in Numerical Integration, T. O. Espelid and A. Genz, eds., Kluwer Acad. Publ., Dordrecht, the Netherlands, 1992, pp. 37–44.

[3] B. FUGLEDE, *The multidimensional moment problem*, Exposition. Math., 1 (1983), pp. 47–65.

[4] G. GASPER AND M. RAHMAN, *Basic hypergeometric series*, Cambridge University Press, Cambridge, U.K., 1990.

[5] H. L. KRALL AND I. M. SHEFFER, *Orthogonal polynomials in two variables*, Ann. Mat. Pura Appl. (4), 76 (1967), pp. 325–376.

[6] H. M. MÖLLER, *Kubaturformeln mit minimaler Knotenzahl*, Numer. Math., 25 (1976), pp. 185–200.

[7] G. G. RASPUTIN, *On the question of numerical characteristics for orthogonal polynomials of two variables*, Metody Vychisl., 13 (1983), pp. 145–154. (In Russian.)

[8] H. J. SCHMID, *Two-dimensional minimal cubature formulas and matrix equations*, SIAM J. Matrix Anal. Appl., to appear.

[9] A. H. STROUD, *Approximate calculation of multiple integrals*, Prentice Hall, Englewood Cliffs, NJ, 1971.

[10] Y. XU, *Multivariate orthogonal polynomials and operator theory*, Trans. Amer. Math. Soc., 343 (1994), pp. 193–202.

[11] ———, *On multivariate orthogonal polynomials*, SIAM J. Math. Anal., 24 (1993), pp. 783–794.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 488–499, March 1995

013

CONVEX APPROXIMATION BY RATIONAL FUNCTIONS*

BO GAO†, DONALD J. NEWMAN†, AND V. A. POPOV‡

This paper is dedicated to Dick Askey and Frank Olver on the occasion of their birthdays.

Abstract. In this paper, the convex approximation to $|x|$ and to convex functions with continuous derivatives are investigated. In the first case, the approximation order $c_1 e^{-c_2\sqrt{n}}$ is achieved by using H^∞ quadrature. In the second case, the estimate $| f(x) - R_n(x) | \leq C\frac{1}{n^{2-\epsilon}}$ is proved, where ϵ is any positive real number.

Key words. convex approximation, rational functions

AMS subject classifications. 41A20, 41A25, 41A29

1. Introduction. The main result of this paper brings a convex approximation by rational functions to a function $f(x) \in \text{Conv}\,[a,\ b]\bigcap C^1[a,\ b]$. The idea for this work comes from a proof to Newman's conjecture (see [3]). Because of its nonlinearity, rational approximation is inherently more difficult than polynomial approximation. The restriction of "form fitting" makes convex rational approximation an even more difficult problem, and very few results are known.

Let R_n denote the set of all rational functions. We first construct a convex approximation to $|x|$ on $[-1, 1]$, and then we extend this result to the whole real line. Since $f(x)$ can be approximated by a polygon, we obtain

$$R_n^*(f) \leq C\,\frac{\|f'\|_{C[a,\,b]}}{n^{2-\epsilon}},$$

where $R_n^*(f)$ is defined by

$$R_n^*(f) = \inf_{\substack{r(x)\in R_n \\ r(x) \text{ is convex}}} \|f(x) - r(x)\|_{C[a,\,b]}.$$

2. Convex approximation to $|x|$ on $[-1, 1]$. The function $|x|$ plays a very important role in approximation theory. As a consequence, rational approximation to $|x|$ has been intensely studied in [2], [4], and [5]. Since the original construction of Newman [2] is in fact a comonotone approximation to $|x|$, it is natural to hope that the same or a modified construction can give a convex approximation to $|x|$. However, this is not the case, because any function that interpolates $|x|$ cannot be convex.

The existence of a convex approximation to $|x|$ comes from the fact that the function $x\arctan(Nx)$ can also give a good approximation to $|x|$ for large N. Although $\arctan(Nx)$ is not a rational function, its integral representation and the theory for super quadrature make it possible to construct a rational function which gives a convex approximation to $|x|$.

From the work of Andersson and Bojanov [1], we have the following theorem.

* Received by the editors June 19, 1992; accepted for publication (in revised form) October 19, 1993. This paper was originally submitted for the special issue dedicated to Frank Olver and Richard Askey (*SIAM J. Math. Anal.*, March 1994, Vol. 25, No. 2).

† Department of Mathematics, Temple University, Philadelphia, PA 19122.

‡ Department of Mathematics, Temple University, Philadelphia, PA 19122, until his death in 1991.

THEOREM A. *For $i = 1, 2, \ldots, n$, there exist $t_i \in [-1, 1]$ and $A_i \geq 0$ such that*

$$\sup_{\|f\|_{H^\infty(D)} \leq 1} \left| \int_{-1}^{1} f(t)dt - \sum_{i=1}^{n} A_i f(t_i) \right| \leq C n e^{-\pi n} \leq C' e^{-3\sqrt{n}},$$

where D is the unit disk and $H^\infty(D)$ is the space of all functions which are analytic and bounded in D.

Obviously this result can be reformulated by conformal mapping in the following way.

THEOREM A′. *There exist $t_i \in (0, 1)$ and $A_i \geq 0$, $i = 1, 2, \ldots, n$, such that*

$$\sup_{\|f\|_{H^\infty(\Omega)} \leq M} \left| \int_{0}^{1} f(t)dt - \sum_{i=1}^{n} A_i f(t_i) \right| \leq C M e^{-3\sqrt{n}},$$

where C is an absolute constant and $\Omega = \{z : |z - \frac{1}{2}| \leq \frac{1}{2}\}$.

Now consider the function $g_N(x) = (2/\pi)\, x \arctan(Nx)$. From elementary calculus it is easy to prove the next lemma.

LEMMA 1. *For $x \geq 0$ and $N > 0$, we have*

$$0 \leq x - g_N(x) = \frac{2x}{\pi} \int_{Nx}^{\infty} \frac{dt}{1 + t^2} \leq \frac{2}{\pi N}, \tag{1}$$

$$g_N(x) = \frac{2}{\pi} \int_{0}^{1} \frac{Nx^2}{1 + N^2 x^2 t^2}\, dt, \tag{2}$$

$$g_N'(x) = \frac{2}{\pi} \left(\arctan(Nx) + \frac{Nx}{1 + N^2 x^2} \right) = \frac{2}{\pi} \int_{0}^{1} \frac{2Nx}{(1 + N^2 x^2 t^2)^2}\, dt, \tag{3}$$

and

$$g_N''(x) = \frac{2}{\pi} \frac{2N}{(1 + N^2 x^2)^2} = \frac{2}{\pi} \int_{0}^{1} \frac{2N(1 - 3t^2 N^2 x^2)}{(1 + t^2 N^2 x^2)^3}\, dt \geq \frac{C_0}{N^3}. \tag{4}$$

Before using the quadrature formula to obtain our theorem, we need the following lemma.

LEMMA 2. *For $a \geq 0$ and $\Omega = \{z : |z - \frac{1}{2}| \leq \frac{1}{2}\}$ we have*

$$\left| \frac{1}{1 + z^2 a^2} \right| \leq 2 \max\{2, a\}.$$

Proof. Let $z = x + iy$. For $|z - \frac{1}{2}| = \frac{1}{2}$, we have

$$y^2 = x - x^2,$$

$$1 + z^2 a^2 = 1 + (2x^2 - x)a^2 + 2ixya^2,$$

and

$$\max_{z \in \Omega} \left| \frac{1}{1 + z^2 a^2} \right| = \frac{1}{\min_{z \in \Omega} |1 + z^2 a^2|},$$

but

$$\min_{z\in\Omega}|1+z^2a^2| = \min_{y^2=x-x^2}|1+(2x^2-x)a^2+2ixya^2|$$
$$\geq \frac{1}{2}\min_{y^2=x-x^2}|1+(2x^2-x)a^2+2x|y|a^2|$$
$$\geq \begin{cases} \frac{1}{4} & \text{if } 0\leq x\leq \frac{1}{2a^2} \quad (\text{since } 1-xa^2+2x^2a^2\geq \frac{1}{2}),\\ \frac{1}{2a} & \text{if } \frac{1}{2a^2}\leq x\leq \frac{1}{2} \quad (\text{since } |y|\geq \sqrt{\frac{1}{2a^2}}\sqrt{1-x}\geq \frac{1}{a}),\\ \frac{1}{2} & \text{if } \frac{1}{2}\leq x\leq 1 \quad (\text{since } 1-xa^2+2x^2a^2\geq 1). \end{cases}$$

From the last inequality, Lemma 2 follows.

Using Lemmas 1 and 2 and Theorem A′ we can prove the following theorem.

THEOREM 1. *There exist a constant C and a rational function $r(x)$ of degree n such that for $x\in[-1,1]$*

$$\text{(a)} \qquad ||x|-r(x)|\leq C\,e^{-\frac{\sqrt{n}}{3}}$$

and

$$\text{(b)} \qquad r''(x)\geq 0.$$

Proof. By applying Lemma 1 and Theorem A′ to (2)–(4), we obtain

$$\left\|\frac{Nx^2}{1+z^2N^2x^2}\right\|_{H^\infty(\Omega)} \leq 2\,N\,x^2\,\max(2,\,Nx)\leq C_1\,N^2,$$

$$\left\|\frac{2Nx}{(1+z^2N^2x^2)^2}\right\|_{H^\infty(\Omega)} \leq 2\,N\,x\,(2\,\max(2,\,Nx))^2\leq C_2\,N^3,$$

and

$$\left\|\frac{2N(1-3z^2N^2x^2)}{(1+z^2N^2x^2)^3}\right\|_{H^\infty(\Omega)} \leq (2N+6\,N^3\,x^2)\,(2\,\max(2,\,Nx))^3\leq C_3\,N^6.$$

These inequalities imply that

$$\text{(5)} \qquad \left|g_N(x)-\frac{2}{\pi}\sum_{i=1}^{n}A_i\,\frac{Nx^2}{1+t_i^2x^2N^2}\right|\leq C_4\,N^2\,e^{-3\sqrt{n}},$$

$$\left|g_N'(x)-\frac{2}{\pi}\sum_{i=1}^{n}A_i\,\frac{2Nx}{(1+t_i^2x^2N^2)^2}\right|\leq C_5\,N^3\,e^{-3\sqrt{n}},$$

and

$$\text{(6)} \qquad \left|g_N''(x)-\frac{2}{\pi}\sum_{i=1}^{n}A_i\,\frac{2N(1-3t_i^2N^2x^2)}{(1+t_i^2N^2x^2)^3}\right|\leq C_6\,N^6\,e^{-3\sqrt{n}},$$

respectively. Let

$$r(x)=\frac{2}{\pi}\sum_{i=1}^{n}A_i\,\frac{Nx^2}{1+t_i^2x^2N^2}.$$

By (5) and (1), we have

$$||x| - r(x)| \le ||x| - g_N(x)| + |g_N(x) - r(x)| \le \frac{2}{\pi N} + C_4 N^2 e^{-3\sqrt{n}}. \tag{7}$$

Pick

$$N = \frac{1}{2}\left(\frac{C_0}{C_6}\right)^{\frac{1}{9}} e^{\frac{\sqrt{n}}{3}}$$

in (7), so that for $x \ge 0$ we have

$$||x| - r(x)| \le \frac{1}{\pi}\left(\frac{C_6}{C_0}\right)^{\frac{1}{9}} e^{-\frac{\sqrt{n}}{3}} + \frac{C_4}{4}\left(\frac{C_0}{C_6}\right)^{\frac{2}{9}} e^{-\frac{7}{3}\sqrt{n}} \le C\, e^{-\frac{\sqrt{n}}{3}}.$$

From

$$\frac{C_6}{C_0} N^9 e^{-3\sqrt{n}} = \left(\frac{1}{2}\right)^9 < 1,$$

we find that

$$C_6 N^6 e^{-3\sqrt{n}} < \frac{C_0}{N^3}.$$

Then by (4) and (6) we see that for $x \ge 0$, $r''(x) > 0$. Since $r(x)$ and $r''(x)$ are even functions, and since we have shown the necessary results for $x \ge 0$, Theorem 1 is proven for $x \in [-1, 1]$.

3. Convex approximation to $f(x) \in \mathbf{Conv}\,[a, b] \bigcap C^1[a, b]$. Now, we consider the convex approximation to $f(x) \in \mathrm{Conv}\,[a, b] \bigcap C^1[a, b]$, the space of all convex functions with continuous derivative. This space was the first function space on which the rational approximation was shown to be better than the polynomial approximation. The main result we have obtained is the following theorem.

THEOREM 2. *For every $\epsilon > 0$, n large and $f(x) \in \mathrm{Conv}\,[a, b] \bigcap C^1[a, b]$, there exists a rational function $r(x)$ of degree n which is convex on the interval $[a, b]$ such that*

$$\|f(x) - r(x)\|_{C[a, b]} \le C \frac{\|f'\|_{C[a, b]}}{n^{2-\epsilon}},$$

where C is an absolute constant.

Before we give the proof of Theorem 2, we need the definition of the ski-slope function and some preliminary results.

DEFINITION 1. *The "ski-slope" function $k(x)$ on $x \in [0, \infty]$ is defined by*

$$k(x) = \begin{cases} 1 - x, & 0 \le x < 1, \\ 0, & x \ge 1. \end{cases}$$

We first prove that any convex polygonal function $g(x)$ on $[0, 1]$ can be written as a linear combination of $k(x)$.

LEMMA 3. *Given the interval $[0, 1]$ and a partition of the interval $0 = x_0 < x_1 < \cdots < x_N = 1$, let $g(x)$ be any convex polygonal function on $[0, 1]$ with vertices*

$x_0, \ldots, x_N$ such that $g(x)$ is a nonincreasing function and $g(1) = 0$. Then there exists a linear combination of $k(x)$, defined by

$$K(x) = \sum_{i=0}^{N-1} a_i\, k\left(\frac{x}{x_{i+1}}\right),$$

with

$$a_{N-1} = \frac{g(x_{N-1})}{1 - x_{N-1}}$$

and

$$a_i = \frac{g(x_i) - \displaystyle\sum_{j=i+1}^{N-1} a_j\, k\left(\frac{x_i}{x_{j+1}}\right)}{1 - \dfrac{x_i}{x_{i+1}}}, \qquad i = N-2,\ \ldots,\ 1,\ 0$$

such that

$$K(x) = g(x) \quad \text{for } x \in [0, 1] \tag{8}$$

and

$$0 \le a_i \le 2\|g'\|_{L^\infty[0,1]}. \tag{9}$$

Proof. By the definition of $k(x)$, for $i = 0,\ 1,\ \ldots,\ N$, we have

$$k\left(\frac{x}{x_{i+1}}\right) = \begin{cases} 1 - \dfrac{x}{x_{i+1}}, & 0 \le x < x_{i+1}, \\ 0, & x \ge x_{i+1}. \end{cases}$$

Therefore, for $x \in [x_i,\ x_{i+1}]$, $i = 0,\ 1,\ \ldots,\ N-1$, we have

$$K(x) = \sum_{j=i}^{N-1} a_j\, k\left(\frac{x}{x_{j+1}}\right) \tag{10}$$

and

$$\begin{aligned} K(x_i) &= a_i\, k\left(\frac{x_i}{x_{i+1}}\right) + \sum_{j=i+1}^{N-1} a_j\, k\left(\frac{x_i}{x_{j+1}}\right) \\ &= \frac{g(x_i) - \displaystyle\sum_{j=i+1}^{N-1} a_j\, k\left(\frac{x_i}{x_{j+1}}\right)}{1 - \dfrac{x_i}{x_{i+1}}}\, k\left(\frac{x_i}{x_{i+1}}\right) + \sum_{j=i+1}^{N-1} a_j\, k\left(\frac{x_i}{x_{j+1}}\right) \\ &= g(x_i) - \sum_{j=i+1}^{N-1} a_j\, k\left(\frac{x_i}{x_{j+1}}\right) + \sum_{j=i+1}^{N-1} a_j\, k\left(\frac{x_i}{x_{j+1}}\right) = g(x_i). \end{aligned}$$

Obviously, $K(x_N) = K(1) = 0 = g(x_N)$. Thus we have proved that (8) holds for $x = x_i$, $i = 0,\ 1,\ \ldots,\ N$. Since $K(x)$ is also a polygonal function with the same

vertices as $g(x)$, we have $K(x) = g(x)$ for $x \in [0, 1]$. Now if we assume that c_i is the slope of $g(x)$ on the interval $[x_i, x_{i+1}]$, then $c_i - c_{i+1} \leq 0$ for $i = 0, \ldots, N-1$, and by (10), we have

$$\sum_{j=i}^{N-1} a_j \, k\left(\frac{x}{x_{j+1}}\right) = g(x) = c_i(x - x_i) + g(x_i) \quad x \in [x_i, x_{i+1}].$$

Since the left-hand side is just a straight line for $[0, x_{i+1}]$, then in fact we have

$$\sum_{j=i}^{N-1} a_j \, k\left(\frac{x}{x_{j+1}}\right) = c_i(x - x_i) + g(x_i) \quad x \in [0, x_{i+1}].$$

Therefore, we obtain for $i = 0, \ldots, N-2$,

$$\begin{aligned} a_i &= \frac{g(x_i) - \sum_{j=i+1}^{N-1} a_j \, k\left(\frac{x_i}{x_{j+1}}\right)}{1 - \frac{x_i}{x_{i+1}}} \\ &= \frac{g(x_i) - (c_{i+1}(x_i - x_{i+1}) + g(x_{i+1}))}{1 - \frac{x_i}{x_{i+1}}} \\ &= x_{i+1}\,(c_{i+1} - c_i) \geq 0. \end{aligned}$$

Finally, since $0 \leq x_{i+1} \leq 1$, for $i = 0, \ldots, N-2$, we have

$$0 \leq a_i \leq 2 \max |c_i| \leq 2\|g'\|_{L^\infty[0,1]}.$$

In addition, these inequalities hold for a_{N-1} by its definition and the mean value theorem. □

LEMMA 4. *For $f(x) \in \mathrm{Conv}[a, b] \cap C^1[a, b]$, there exists a partition of $[a, b]$, $a = x_0 < x_1 < \cdots < x_N = b$ such that the polygonal function $g(x)$, which has the x_i's as its vertices and interpolates $f(x)$ at the x_i's, satisfies*

$$\|f(x) - g(x)\|_{C[a,b]} \leq 2(b-a)\frac{\|f'\|_{C[a,b]}}{N^2}. \tag{11}$$

Proof. Without loss of generality, we can assume that $[a, b] = [0, 1]$. Let

$$F(x) = \frac{f(x) - f'(0)x}{|f'(1) - f'(0)|}.$$

Then we have that $F(x) \in \mathrm{Conv}[0,1] \cap C^1[0,1], F'(0) = 0, F'(x) \geq 0$, and $F'(1) = \|F'(x)\|_{C[0,1]} = 1$. Since $x + F'(x)$ is an increasing function on $[0, 1]$, the x_i's which satisfy $x_i + F'(x_i) = (2i/N)$ for $i = 0, 1, \ldots, N$ form a partition of [0,1]. Now let $g_1(x)$ be the polygonal function with the x_i's as its vertices, and let $g_1(x)$ interpolate $F(x)$ at the x_i's. For $x \in [x_{i-1}, x_i]$ we have

$$\begin{aligned} |F(x) - g_1(x)| &\leq (x_i - x_{i-1})\,[F'(x_i) - F'(x_{i-1})] \\ &\leq \frac{1}{4}\,[x_i - x_{i-1} + F'(x_i) - F'(x_{i-1})]^2 \\ &= \frac{1}{4}\left(\frac{2}{N}\right)^2 = \frac{1}{N^2}. \end{aligned}$$

Now let $g(x) = |f'(1)-f'(0)|\, g_1(x)+f'(0)x$ to have (11) hold for $[a, b] = [0, 1]$. Finally, for a general closed interval $[a, b]$, (11) holds by a linear transformation. □

COROLLARY 1. *If $N \geq 2$, $f(x) \in \mathrm{Conv}[0,1] \cap C^1[0,1]$, $f'(x) \leq 0$ and $f(1) = 0$, then there exists a partition of $[0,1]$, $0 = x_0 < x_1 < \cdots < x_N = 1$, and a function $K(x)$ as defined in Lemma 3 such that*

$$x_1 = \frac{1}{N^2} \tag{12}$$

and

$$\|f(x) - K(x)\|_{C[0,1]} \leq 6\frac{\|f'(x)\|_{C[0,1]}}{N^2}\,. \tag{13}$$

Proof. The partition of [0,1] is obtained by setting $x_1 = \frac{1}{N^2}$ and applying Lemma 4 to $f(x)$ on the interval $[x_1, 1]$ with N replaced by $N-1$. Let $g(x)$ be the resulting polygonal function with $g(0) = f(0)$. We then have

$$|f(x) - g(x)| \leq [x_1 - 0][f'(x_1) - f'(0)] \leq 2\frac{\|f'(x)\|_{C[0,1]}}{N^2} \quad \text{for } x \in [0, x_1]$$

and

$$|f(x) - g(x)| \leq 2\left(1 - \frac{1}{N^2}\right)\frac{\|f'(x)\|_{C[0,1]}}{(N-1)^2} \leq 6\frac{\|f'(x)\|_{C[0,1]}}{N^2} \quad \text{for } x \in [x_1, 1].$$

Corollary 1 then follows from Lemma 3. □

Now we extend the results of §2 into the whole real line.

LEMMA 5. *Let ν be a positive integer and let $\mu = e^{\sqrt{\nu}/3}$. Define the rational function*

$$r(x) = \frac{2}{\pi}\sum_{i=1}^{\nu} A_i \frac{\mu x^2}{1 + t_i^2 x^2 \mu^2},$$

where the A_i's and the t_i's are constants determined by Theorem 1. Then for $x \in [-1, 1]$,

$$\||x| - r(x)\| \leq \frac{C}{\mu}, \tag{14}$$

$$r'(x) = \frac{2}{\pi}\left(\arctan(\mu x) + \frac{\mu x}{1 + x^2\mu^2}\right) + O\left(\frac{1}{\mu^3}\right), \tag{15}$$

and

$$r''(x) = \frac{2}{\pi}\left(\frac{\mu}{(1 + x^2\mu^2)^2}\right) + O\left(\frac{1}{\mu^3}\right). \tag{16}$$

In addition, for $|x| \geq 1$ we have

$$|r(x)| \leq C\, x^2, \tag{17}$$

$$|r'(x)| \leq 2C\, |x|, \tag{18}$$

and

$$|r''(x)| \le 8C, \tag{19}$$

where C is a constant.

Proof. (14)–(16) are direct results of Theorem 1, and by the definition of $r(x)$, for $x \ge 1$, we have

$$|r'(x)| = \frac{2}{\pi}\sum_{i=1}^{\nu} A_i \frac{2\mu x}{(1+t_i^2\mu^2x^2)^2} \le \frac{2r(x)}{x}$$

and

$$|r''(x)| = \left|\frac{2}{\pi}\sum_{i=1}^{\nu} A_i \frac{2\mu(1-3t_i^2\mu^2x^2)}{(1+t_i^2\mu^2x^2)^3}\right| \le \frac{r'(x)}{x} + 3\frac{r'(x)}{x} = 4\frac{r'(x)}{x}.$$

Solving these two differential inequalities with the initial condition $r(1) = 1+O(e^{-\sqrt{\nu}/3})$, we find that (17)–(19) hold for $x \ge 1$. These results also hold for $x \le -1$ by the symmetric property of $r(x)$. □

From Lemma 5 we have the next theorem.

THEOREM 3. *Let ν be a positive integer and set $\mu = e^{\sqrt{\nu}/3}$. Define $r(x)$ to be the function in Lemma 5 and let $T(x) = \frac{1}{2}(|x-1| - 2|x| + |x+1|)$. Also define*

$$S(x) = \frac{3}{2}\left[r\left(\frac{x-1}{3}\right) - 2r\left(\frac{x}{3}\right) + r\left(\frac{x+1}{3}\right)\right], \tag{20}$$

$$\eta(x) = \left(1 + \left(\frac{x^2}{2}\right)^{\nu}\right)^{-1}, \tag{21}$$

$$\eta_1(x) = \mu^{\frac{3}{2}} \frac{x^2}{(\mu^2+x^2)^2}, \tag{22}$$

and

$$t(x) = S(x)\,\eta(x) + \eta_1(x).$$

For large ν we then have

$$3r\left(\frac{x}{3}\right) = |x| + O\left(\frac{1}{\mu}\right) \quad \text{for } x \in [-3,3], \tag{23}$$

$$|T(x) - S(x)| \le O\left(\frac{1}{\mu}\right) \quad \text{for } x \in [-2,2], \tag{24}$$

$$|T(x) - t(x)| \le \frac{C_1}{\mu} \quad \text{for } x \in (-\sqrt{\mu},\ \sqrt{\mu}), \tag{25}$$

$$t''(x) \ge 0 \quad \text{for } x \in \left[\tfrac{1}{2},\ \sqrt{\mu}\right], \tag{26}$$

and

$$\text{the degree of } t(x) \le 10\nu\,, \tag{27}$$

where C_1 is an absolute constant.

Proof. (23) is evident from Theorem 1. (24) and (27) can be seen easily after some simple calculations. To prove (25), we first obtain

$$\eta(x) = \begin{cases} 1 + O\left(\frac{1}{2^\nu}\right), & 0 \le |x| \le 1, \\ 1 + O(1), & 1 \le |x| \le 2, \\ O\left(\frac{1}{2^{\frac{\nu}{2}}}\right)\left(\frac{1}{(1+x^2)^2}\right), & |x| \ge 2, \end{cases} \tag{28}$$

and

$$|\eta_1(x)| \le O\left(\frac{1}{\mu^{\frac{3}{2}}}\right) \quad \text{for } x \in [-\sqrt{\mu}, \sqrt{\mu}], \tag{29}$$

from (21) and (22), respectively. Now by (23) and (24), for $0 \le |x| \le 1$,

$$\begin{aligned} |T(x) - t(x)| &\le |T(x) - S(x)\eta(x)| + \frac{1}{\mu^{\frac{3}{2}}} \\ &\le |T(x) - S(x)| + |S(x)| O\left(\frac{1}{2^\nu}\right) + \frac{1}{\mu^{\frac{3}{2}}} \\ &\le O\left(\frac{1}{\mu}\right) + O\left(\frac{1}{2^\nu}\right) + \left(\frac{1}{\mu^{\frac{3}{2}}}\right) = O\left(\frac{1}{\mu}\right). \end{aligned} \tag{30}$$

Note that for $|x| \ge 1$, $T(x) = 0$. Because of this fact and by (28), (29), and (24), for $1 \le |x| \le 2$, we have

$$\begin{aligned} |T(x) - t(x)| &\le |T(x) - S(x)\eta(x)| + \frac{1}{\mu^{\frac{3}{2}}} \\ &\le |T(x) - S(x)||\eta(x)| + \frac{1}{\mu^{\frac{3}{2}}} \\ &\le O\left(\frac{1}{\mu}\right) + \frac{1}{\mu^{\frac{3}{2}}} = O\left(\frac{1}{\mu}\right). \end{aligned} \tag{31}$$

For $2 \le |x| \le \sqrt{\mu}$, by (17), (28), and (29) together with $T(x) = 0$, we find

$$\begin{aligned} |T(x) - t(x)| &\le |S(x)\eta(x)| + \frac{1}{\mu^{\frac{3}{2}}} \\ &\le O\left(\frac{1}{2^\nu}\frac{x^2}{(1+x^2)^2}\right) + \frac{1}{\mu^{\frac{3}{2}}} \le O\left(\frac{1}{\mu}\right). \end{aligned} \tag{32}$$

Then, (25) follows directly from (30)–(32).

For (26), and from (21) and (22), we first compute

$$\eta'(x) = \frac{-\nu x\left(\frac{x^2}{2}\right)^{\nu-1}}{\left(1 + \left(\frac{x^2}{2}\right)^\nu\right)^2}, \tag{33}$$

$$\eta''(x) = \frac{\nu(2\nu+1)\left(\frac{x^2}{2}\right)^{\nu-1}\left(\left(\frac{x^2}{2}\right)^{\nu} - \frac{2\nu-1}{2\nu+1}\right)}{\left(1+\left(\frac{x^2}{2}\right)^{\nu}\right)^3}, \tag{34}$$

and for $x \in [-\sqrt{\mu}, \sqrt{\mu}]$,

$$\eta_1''(x) = \frac{\mu^{\frac{3}{2}}(3x^4 - 8x^2\mu^2 + \mu^4)}{(\mu^2+x^2)^4} \geq \frac{1}{\mu^{\frac{5}{2}}}. \tag{35}$$

Furthermore, let $x_0 = \sqrt{2}\left(\frac{2\nu-1}{2\nu+1}\right)^{1/(2\nu)}$ be the solution of $\eta''(x) = 0$. Then by (33) and (34), there exists $\delta \in (0, \frac{1}{3}]$ such that

$$\max(|\eta(x)|,\ |\eta'(x)|,\ |\eta''(x)|) \leq \frac{C_2\nu^2}{(1+\delta)^{\nu}(1+x^2)}, \quad x \in [x_0+\delta, \sqrt{\mu}], \tag{36}$$

$$\max(|\eta'(x)|,\ |\eta''(x)|) \leq \frac{C_2\nu^2}{(1+\delta)^{\nu}}, \quad x \in \left[\frac{1}{2}, x_0-\delta\right], \tag{37}$$

and

$$\max(|\eta'(x)|,\ |\eta''(x)|) \leq C_3\nu^2, \quad x \in [x_0-\delta,\ x_0+\delta]. \tag{38}$$

Now from (17)–(19), (35), and (36), together with

$$t''(x) = \eta_1''(x) + S''(x)\eta(x) + 2S'(x)\eta'(x) + S(x)\eta''(x) \tag{39}$$

and

$$\lim_{\nu\to\infty} \frac{\mu^k}{(1+\delta)^{\nu}} = 0 \qquad \text{for } k > 0\,, \tag{40}$$

we obtain

$$t''(x) \geq \frac{1}{\mu^{\frac{5}{2}}} - (64 + 16x + 8x^2)\frac{C\,C_2\,\nu^2}{(1+\delta)^{\nu}(1+x)^2} \geq 0$$

for $x_0 + \delta \leq x \leq \sqrt{\mu}$ and ν large.

For $x \in [\frac{1}{2},\ x_0+\delta]$, by the definition of $S(x)$ and (16), we have

$$\begin{aligned} S''(x) &= \frac{3}{2}\left[r''\left(\frac{x-1}{3}\right) - 2r''\left(\frac{x}{3}\right) + r''\left(\frac{x+1}{3}\right)\right], \\ &= \frac{3}{\pi}\left(\frac{\mu}{(1+(\frac{x-1}{3})^2\mu^2)^2} - 2\frac{\mu}{(1+(\frac{x}{3})^2\mu^2)^2} + \frac{\mu}{(1+(\frac{x+1}{3})^2\mu^2)^2}\right) + O\left(\frac{1}{\mu^3}\right) \\ &\geq -\frac{C_4}{\mu^3}, \qquad x \in \left[\frac{1}{2},\ x_0+\delta\right]. \end{aligned} \tag{41}$$

For $x \in [\frac{1}{2},\ x_0-\delta]$, from (14), (15), and the definition of $S(x)$, it is easy to obtain $|S(x)| \leq 12$ and $|S'(x)| \leq 6\max_{-1\leq x\leq 1}|r'(x)| \leq 12$; combining those facts, $0 \leq \eta(x) \leq 1$, (37), (39) and (41), we find that

$$t''(x) \geq \frac{1}{\mu^{\frac{5}{2}}} - \frac{C_4}{\mu^3}\eta(x) - \frac{12C_2\,\nu^2}{(1+\delta)^{\nu}} - \frac{12\,C_2\,\nu^2}{(1+\delta)^{\nu}} \geq 0, \quad x \in \left[\frac{1}{2},\ x_0-\delta\right]$$

holds for ν large.

Finally, for $x \in (x_0 - \delta, x_0 + \delta)$, from the definition of $S(x)$ and mean value theorem, we have

$$S(x) = \frac{3}{2} \cdot \frac{2}{9} r''(\xi_1), \quad \xi_1 \in \left(\frac{x_0 - \delta - 1}{3}, \frac{x_0 + \delta + 1}{3} \right).$$

By the fact that $\frac{x_0 - \delta - 1}{3} > 0$, $\frac{x_0 + \delta + 1}{3} < 1$, and (16), we obtain

$$|S(x)| \leq \frac{C_5}{\mu^3}, \qquad x \in [x_0 - \delta, x_0 + \delta]. \tag{42}$$

Similarly, if we replace $r'(x)$ in $S'(x)$ by (15) and from that

$$\left(\frac{2}{\pi} \left(\arctan(\mu x) + \frac{\mu x}{1 + x^2 \mu^2} \right) \right)'' = \frac{2}{\pi} \frac{-4x\mu^3}{(1 + x^2\mu^2)^3},$$

we have

$$S'(x) = \frac{2}{3\pi} \frac{-4\xi_2 \mu^3}{(1 + \xi_2^2 \mu^2)^3} + O\left(\frac{1}{\mu^3} \right), \quad \xi_2 \in \left(\frac{x_0 - \delta - 1}{3}, \frac{x_0 + \delta + 1}{3} \right)$$

and

$$|S'(x)| \leq \frac{C_6}{\mu^3}, \qquad x \in [x_0 - \delta, x_0 + \delta]. \tag{43}$$

From (41)–(43), (35), (38), and (39), we have that

$$t''(x) \geq \frac{1}{\mu^{\frac{5}{2}}} - \frac{C_4}{\mu^3} \eta(x) - \frac{C_5}{\mu^3} C_3 \nu^2 - \frac{C_6}{\mu^3} C_3 \nu^2 \geq 0, \qquad x \in [x_0 - \delta,\ x_0 + \delta]$$

holds for ν large. Therefore, we have shown that $t''(x) \geq 0$ for $x \in [\frac{1}{2}, \sqrt{\mu}]$. □

By Theorem 3 and the fact that $k(x) = 2T(\frac{x+1}{2})$ for $x \geq 0$, we have the following corollary.

COROLLARY 2. *For $x \in [0, \sqrt{\mu}]$, let*

$$\tilde{t}(x) = 2\, t\left(\frac{x+1}{2} \right) = 2 \left[S\left(\frac{x+1}{2} \right) \eta\left(\frac{x+1}{2} \right) + \eta_1\left(\frac{x+1}{2} \right) \right].$$

Then we have

$$\|k(x) - \tilde{t}(x)\| \leq \frac{C}{\mu} \tag{44}$$

and

$$\tilde{t}''(x) \geq 0\,. \tag{45}$$

We are now ready to demonstrate the proof of Theorem 2.

Proof. Without loss of generality, we set $[a, b] = [0, 1]$ and for $f(x) \in \text{Conv}\,[0, 1] \bigcap C^1[0, 1]$, we assume that $f'(x) \leq 0$, $\|f'\|_{C[0,1]} = 1$, and $f(1) = 0$. By Corollary 1, there exist x_i and $a_i, i = 0, 1, \ldots, N-1$ such that

$$\|f(x) - \sum_{i=0}^{N-1} a_i k\left(\frac{x}{x_{i+1}} \right) \|_{C[0,1]} \leq \frac{6}{N^2}.$$

If we set

$$R(x) = \sum_{i=0}^{N-1} a_i \tilde{t}\left(\frac{x}{x_{i+1}}\right)$$

by (12), we have $(x/x_i) \leq N^2$ for $x \in [0, 1]$. If $N^2 \leq \sqrt{\mu}$, from (44), (45), and $a_i \geq 0$, we obtain

$$R''(x) \geq 0 \qquad \text{for } x \in [0, 1],$$

and

$$\begin{aligned} |f(x) - R(x)| &\leq \left| f(x) - \sum_{i=0}^{N-1} a_i\, k\left(\frac{x}{x_{i+1}}\right)\right| \\ &\quad + \left|\sum_{i=0}^{N-1} a_i\, k\left(\frac{x}{x_{i+1}}\right) - \sum_{i=0}^{N-1} a_i\, \tilde{t}\left(\frac{x}{x_{i+1}}\right)\right| \\ &\leq \frac{6}{N^2} + N\frac{C}{\mu}\,. \end{aligned} \tag{46}$$

Now given $\epsilon > 0$, let $N = n^{1-\epsilon}$ and $\nu = n^{\epsilon}$. For large n we have $N^2 \leq \sqrt{\mu}$ and $N^3 \ll \mu$. By (46), and because the degree of $R(x)$ is $10n$, we finally get Theorem 2.

REFERENCES

[1] J. E. ANDERSSON AND B. D. BOJANOV *A note on the optimal quadrature in H_p*, Numer. Math., 44 (1984), pp. 301–308.

[2] D. J. NEWMAN, *Rational approximation to $|x|$*, Michigan Math. J., 11 (1964), pp. 11–14.

[3] ———*Approximation with rational functions*, CBMS No. 41, American Mathematical Society, Providence, RI, 1979.

[4] P. P. PETRUSHEV AND V. A. POPOV, *Rational approximation of real functions*, Cambridge Univ. Press, Cambridge, UK, 1987

[5] N. S. VJACHESLAVOV, *On the uniform approximation of $|x|$ by rational functions*, Dokl. Akad. Nauk SSSR, 220 (1975), pp. 512–515.

SIAM J. MATH. ANAL.
Vol. 26, No. 2, pp. 500–517, March 1995

014

SPLINE WAVELETS OF SMALL SUPPORT*

DEBAO CHEN†

Abstract. Every mth order cardinal spline wavelet is a linear combination of the functions $\{N_{m+l}^{(l)}(2x-j), j \in \mathbf{Z}\}$. Here the function N_m is the mth order cardinal B-spline. This paper proves that the single function $N_{m+l}^{(l)}(2x)$, or $N_{m+l}^{(l)}(2x-1)$ is a wavelet when m and l satisfy some mild conditions. As l decreases, so does the support of the wavelet. When l increases,the smoothness of the dual wavelet improves. Each wavelet is constructed by spline multiresolution analysis. The dual multiresolution analyses are given.

Key words. wavelet, dual wavelet, wavelet basis, biorthogonal wavelet basis, B-spline, Riesz basis

AMS subject classifications. primary, 41A15, 42C15; secondary 41A05, 41A30, 41A58

1. Introduction.

The simplest example of an orthonormal spline wavelet basis is the Haar basis. The orthonormal spline wavelet bases of higher-order spline wavelets were given by Battle [2] and Lemarié [20] by using different methods. Cohen, Daubechies, and Feauveau constructed biorthogonal wavelet bases of compactly supported wavelets [11], [13]. The most important advantage of Cohen, Daubechies, and Feauveau's construction is that both wavelet and dual wavelet are compactly supported, and are still symmetric or antisymmetric. In particular, they constructed compactly supported spline wavelets with compactly supported dual wavelets. As they pointed out, their theory also can be used to construct noncompactly supported wavelets. Their contributions have been very significant. Chui and Wang [8], [9] introduced the following mth order compactly supported cardinal spline wavelet:

$$\psi_m(x) = \frac{1}{2^{m-1}} \sum_{j=0}^{2m-2} (-1)^j N_{2m}(j+1) N_{2m}^{(m)}(2x-j).$$

Here the cardinal B-splines N_m are defined recursively by the equations

$$N_1(x) = \chi_{[0,1)}(x),$$

$$N_m(x) = (N_{m-1} * N_1)(x) = \int_0^1 N_{m-1}(x-t)\,dt, \qquad m = 2, 3, \ldots .$$

This cardinal spline wavelet has been studied by several other authors as well. (See, for example, Auscher [1], Micchelli [24], and Unser and Aldroubi [27].) The advantage of the cardinal spline wavelet is that the wavelet spaces are kept orthogonal and the wavelets are still symmetric or antisymmetric. The dual wavelet is still an mth-order spline function. In a previous paper [3], we extended Chui and Wang's work and proved that the functions

$$\psi_{m,l;c}(x) = \frac{1}{2^{l-1}} \sum_{j=-1}^{m+l-1} (-1)^j N_{m+l}(j+1+c) N_{m+l}^{(l)}(2x-j), \qquad -1 < c < 1$$

* Received by the editors March 10, 1993; accepted for publication (in revised form) October 14, 1993.

† Department of Mathematics, Oklahoma State University, Stillwater, Oklahoma 74078.

are also mth-order compactly supported spline wavelets; here $m+l$ is even and $l \geq 1$ (if $l = 1$, we assume $-2/3 \leq c \leq 2/3$). When $c = 0$, the wavelet $\psi_{m,l;0} = \psi_{m,l}$ is symmetric or antisymmetric. The dual wavelet $\widetilde{\psi}_{m,l;0} = \widetilde{\psi}_{m,l}$ is an lth-order spline function.

In fact, every mth-order cardinal spline wavelet can be written as

$$\psi_{a,m,l}(x) = \frac{1}{2^{l-1}} \sum_j a_j N_{m+l}^{(l)}(2x - j),$$

where $a = \{a_j\}$ is a real sequence that is either finite or infinite.

We ask the following question: Is the single function $N_{m+l}^{(l)}(2x - j)$ a wavelet? In this paper we prove that it is a wavelet when m and l satisfy some mild conditions. Some proofs in this paper are borrowed from Cohen, Daubechies, and Feauveau [10], [11], [13] and Chui and Wang [7], [8], [9].

Using the derivative of certain functions to construct wavelets is a typical method in the construction of wavelet frames and dyadic wavelet transforms [16], [22]. In this paper, we show that this method is also a source of wavelet bases.

First we define the term "wavelet." For a given function f, we will use throughout this paper the notation

$$f_{j,k}(x) = 2^{j/2} f(2^j x - k), \qquad j, k \in \mathbf{Z}.$$

DEFINITION 1. *An element ψ of $L^2(\mathbf{R})$ is called a "wavelet" if*

(1) *$\{\psi_{j,k}\}_{j,k\in\mathbf{Z}}$ is a Riesz basis for $L^2(\mathbf{R})$, and*

(2) *its dual basis is of the form $\{\widetilde{\psi}_{j,k}\}_{j,k\in\mathbf{Z}}$ for some function $\widetilde{\psi}$ in $L^2(\mathbf{R})$.*

We also call $\widetilde{\psi}$ the dual wavelet of ψ. Sometimes we call ψ and $\widetilde{\psi}$ dual wavelets.

Following Mallat [21] and Meyer [23] we define the multiresolution analysis.

DEFINITION 2. *A multiresolution analysis (MRA) of $L^2(\mathbf{R})$ is a sequence $\{V_j\}_{j\in\mathbf{Z}}$ of closed subspaces of $L^2(\mathbf{R})$ such that the following hold:*

(1) *$V_j \subset V_{j+1}$ for all $j \in \mathbf{Z}$;*

(2) *$\cup_{j=-\infty}^{\infty} V_j$ is dense in $L^2(\mathbf{R})$ and $\cap_{j=-\infty}^{\infty} V_j = \{0\}$;*

(3) *$f(x) \in V_j \iff f(2x) \in V_{j+1}$ for all $j \in \mathbf{Z}$;*

(4) *$f(x) \in V_0 \iff f(x-k) \in V_0$ for all $k \in \mathbf{Z}$;*

(5) *there exists a function $\phi \in V_0$ such that $\{\phi(x-k) : k \in \mathbf{Z}\}$ is a Riesz basis of V_0.*

It is well known [4], [17], [21], [23] that the B-spline N_m can serve as the scaling function ϕ in Definition 2 and generates the mth-order spline MRA $\{V_j^m\}_{j\in\mathbf{Z}}$.

Throughout this paper we use the following convention for the Fourier transform:

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-\omega t i}\, dt. \tag{1.1}$$

The Fourier transform of the scaling function N_m is

$$\widehat{N}_m(\omega) = \left(\frac{1 - e^{-\omega i}}{i\omega}\right)^m. \tag{1.2}$$

Hence, we have

$$\widehat{N}_m(\omega) = P_m(e^{-\frac{\omega}{2}i}) \widehat{N}_m\left(\frac{\omega}{2}\right) = \prod_{k=1}^{\infty} P_m(e^{-2^{-k}\omega i}), \tag{1.3}$$

where

$$P_m(z) = \left(\frac{1+z}{2}\right)^m.$$

2. Preliminaries. Let m_0, $\widetilde{m}_0$, m_1, and $\widetilde{m}_1$ be 2π-periodic functions that satisfy

$$m_0(0) = \widetilde{m}_0(0) = 1, \qquad m_0(\pi) = \widetilde{m}_0(\pi) = 0, \tag{2.1}$$

$$\Delta(\omega) = m_0(\omega)m_1(\omega+\pi) - m_1(\omega)m_0(\omega+\pi) \neq 0, \tag{2.2}$$

and

$$\widetilde{m}_0(\omega) = \overline{m_1(\omega+\pi)}/\overline{\Delta(\omega)}, \qquad \widetilde{m}_1(\omega) = -\overline{m_0(\omega+\pi)}/\overline{\Delta(\omega)}. \tag{2.3}$$

These assumptions lead to

$$m_0(\omega)\overline{\widetilde{m}_0(\omega)} + m_1(\omega)\overline{\widetilde{m}_1(\omega)} = 1, \tag{2.4}$$

$$m_0(\omega)\overline{\widetilde{m}_0(\omega+\pi)} + m_1(\omega)\overline{\widetilde{m}_1(\omega+\pi)} = 0, \tag{2.5}$$

$$m_0(\omega)\overline{\widetilde{m}_0(\omega)} + m_0(\omega+\pi)\overline{\widetilde{m}_0(\omega+\pi)} = 1. \tag{2.6}$$

Also, we suppose that the Fourier coefficients of m_0 and $\widetilde{m}_0$ have exponential decay, or at least belong to l^1. We define, first in the sense of tempered distributions, the scaling functions and wavelets by

$$\begin{aligned}
\widehat{\phi}(\omega) &= \prod_{k=1}^{\infty} m_0(2^{-k}\omega), \\
\widehat{\psi}(\omega) &= m_1\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right), \\
\widehat{\widetilde{\phi}}(\omega) &= \prod_{k=1}^{\infty} \widetilde{m}_0(2^{-k}\omega), \\
\widehat{\widetilde{\psi}}(\omega) &= \widetilde{m}_1\left(\frac{\omega}{2}\right)\widehat{\widetilde{\phi}}\left(\frac{\omega}{2}\right).
\end{aligned}$$

Since we have assumed that m_0 and $\widetilde{m}_0$ vanish at $\omega = \pi$, we can express these filters in a factored form

$$\begin{aligned}
m_0(\omega) &= \left(\frac{1+e^{-i\omega}}{2}\right)^L p(\omega), \\
\widetilde{m}_0(\omega) &= \left(\frac{1+e^{-i\omega}}{2}\right)^{\widetilde{L}} \widetilde{p}(\omega).
\end{aligned}$$

The following theorem was proved in [11] and [13]. The authors of [11] and [13] mainly consider the case in which there are finitely many terms in the Fourier series of m_0, $\widetilde{m}_0$, m_1, and $\widetilde{m}_1$. But as the authors pointed out, the following theorem is still true if there are infinitely many terms, having sufficiently decay.

THEOREM 1. *Suppose that* $[-\pi,\pi] = D_1 \cup D_2 \cup \cdots \cup D_j = \widetilde{D}_1 \cup \widetilde{D}_2 \cup \cdots \cup \widetilde{D}_k$,

and that there exist $q > 0$, $\widetilde{q} > 0$ *so that*

$$\begin{aligned} |p(\omega)| &\le q, \quad \omega \in D_1, & |\widetilde{p}(\omega)| &\le \widetilde{q}, \quad \omega \in \widetilde{D}_1, \\ |p(\omega)p(2\omega)| &\le q^2, \quad \omega \in D_2, & |\widetilde{p}(\omega)\widetilde{p}(2\omega)| &\le \widetilde{q}^2, \quad \omega \in \widetilde{D}_2, \\ &\vdots \qquad \vdots & &\vdots \qquad \vdots \\ |p(\omega)p(2\omega)\cdots p(2^{j-1}\omega)| &\le q^j, \quad \omega \in D_j, & |\widetilde{p}(\omega)\widetilde{p}(2\omega)\cdots\widetilde{p}(2^{k-1}\omega)| &\le \widetilde{q}^k, \quad \omega \in \widetilde{D}_k. \end{aligned}$$

Then

$$|\widehat{\phi}(\omega)| \le C(1+|\omega|)^{-L+K}, \qquad |\widehat{\widetilde{\phi}}(\omega)| \le C(1+|\omega|)^{-\widetilde{L}+\widetilde{K}}$$

with $K = \log_2 q$ *and* $\widetilde{K} = \log_2 \widetilde{q}$. *If* $L - K > 1/2$ *and* $\widetilde{L} - \widetilde{K} > 1/2$, *then the functions* $\widehat{\phi}$, $\widehat{\widetilde{\phi}}$, $\widehat{\psi}$, *and* $\widehat{\widetilde{\psi}}$ *are square integrable so that we can define the functions* ϕ, $\widetilde{\phi}$, ψ, *and* $\widetilde{\psi}$ *via inverse Fourier transforms. Furthermore we have the following items:*

1. *The functions* ϕ *and* $\widetilde{\phi}$ *are dual scaling functions, and they generate dual MRAs.*

2. *The functions* ψ *and* $\widetilde{\psi}$ *are dual wavelets in the sense of Definition* 1.

3. *If* $L - K > 1$, $\widetilde{L} - \widetilde{K} > 1$, *and* K *and* $\widetilde{K}$ *are not integers, then* ϕ, $\psi \in C^{L-K-1}$, *and* $\widetilde{\phi}$, $\widetilde{\psi} \in C^{\widetilde{L}-\widetilde{K}-1}$.

Let us consider any l^1 sequences $\{p_k\}$, $\{q_k\}$ and their associated Laurent series

$$P(z) = \frac{1}{2}\sum_{k=-\infty}^{\infty} p_k z^k, \qquad Q(z) = \frac{1}{2}\sum_{k=-\infty}^{\infty} q_k z^k,$$

which satisfy $P(1) = 1$, $P(-1) = Q(1) = 0$. Following Chui [4] we consider the determinant

$$\Delta_{P,Q}(z) = \begin{vmatrix} P(z) & Q(z) \\ P(-z) & Q(-z) \end{vmatrix}.$$

If $\Delta_{P,Q}(z) \neq 0$ on $|z| = 1$, we define

$$\begin{aligned} G(z) &= \frac{Q(-z)}{\Delta_{P,Q}(z)} = \frac{1}{2}\sum_{n=-\infty}^{\infty} g_n z^n, \\ H(z) &= \frac{-P(-z)}{\Delta_{P,Q}(z)} = \frac{1}{2}\sum_{n=-\infty}^{\infty} h_n z^n. \end{aligned} \tag{2.7}$$

If we let

$$m_0(\omega) = P(e^{-\omega i}), \quad \widetilde{m}_0(\omega) = \overline{G(e^{-\omega i})}, \quad m_1(\omega) = Q(e^{-\omega i}), \quad \widetilde{m}_1(\omega) = \overline{H(e^{-\omega i})}$$

then the conditions (2.1)–(2.3) are satisfied.

3. Main results. In this section, we construct a special kind of biorthogonal wavelet basis. Our wavelets are the B-spline's derivatives of certain orders which are dilated and, in some cases, translated. One advantage of our wavelets is that the wavelets are quite simple and have relatively small supports. We define

$$\xi_{m,l}(x) = \begin{cases} 2^{1-l}N_{m+l}^{(l)}(2x) & \text{if } m+l \text{ is odd, or } (m+l)/2 \text{ is odd,} \\ 2^{1-l}N_{m+l}^{(l)}(2x-1) & \text{if } (m+l)/2 \text{ is even.} \end{cases} \tag{3.1}$$

LEMMA 1. *Let* $z = e^{-\frac{1}{2}\omega i}$. *Then*

$$\widehat{\xi}_{m,l}(\omega) = Q_{m,l}(z)\widehat{N}_m\left(\frac{\omega}{2}\right),$$

where

$$Q_{m,l}(z) = \begin{cases} 2^{-l}(1-z)^l & \text{if } m+l \text{ is odd or } (m+l)/2 \text{ is odd,} \\ 2^{-l}z(1-z)^l & \text{if } (m+l)/2 \text{ is even.} \end{cases} \tag{3.2}$$

Proof. We prove only the first case. For the second case, the proof is similar. Recall that

$$N_{m+l}^{(l)}(x) = \sum_{k=0}^{l}(-1)^k\binom{l}{k}N_m(x-k).$$

When $m+l$ is odd or $(m+l)/2$ is odd, we have

$$\xi_{m,l}(x) = 2^{1-l}\sum_{k=0}^{l}(-1)^k\binom{l}{k}N_m(2x-k).$$

Hence,

$$\widehat{\xi}_{m,l}(\omega) = 2^{-l}\sum_{k=0}^{l}(-1)^k\binom{l}{k}z^k\widehat{N}_m\left(\frac{\omega}{2}\right) = 2^{-l}(1-z)^l\widehat{N}_m\left(\frac{\omega}{2}\right). \qquad \square$$

When $m+l$ is odd or $(m+l)/2$ is odd, we have

$$\Delta_{P_m,Q_{m,l}}(z) = \begin{vmatrix} 2^{-m}(1+z)^m & 2^{-l}(1-z)^l \\ 2^{-m}(1-z)^m & 2^{-l}(1+z)^l \end{vmatrix} = 2^{-(m+l)}\big((1+z)^{m+l} - (1-z)^{m+l}\big).$$

When $(m+l)/2$ is even, we have

$$\Delta_{P_m,Q_{m,l}}(z) = \begin{vmatrix} 2^{-m}(1+z)^m & 2^{-l}z(1-z)^l \\ 2^{-m}(1-z)^m & -2^{-l}z(1+z)^l \end{vmatrix} = -2^{-(m+l)}z\big((1+z)^{m+l} + (1-z)^{m+l}\big).$$

Let $z = e^{-\omega i}$. When $m+l$ is even, whether $(m+l)/2$ is even or odd, we have

$$|\Delta_{P_m,Q_{m,l}}(z)| = \cos^{m+l}\frac{\omega}{2} + \sin^{m+l}\frac{\omega}{2} \neq 0.$$

When $m+l$ is odd, we have

$$\begin{aligned}\Delta_{P_m,Q_{m,l}}(z) &= \left(\frac{1+e^{-\omega i}}{2}\right)^{m+l} - \left(\frac{1-e^{-\omega i}}{2}\right)^{m+l} \\ &= e^{-(m+l)\omega i/2}\left(\cos^{m+l}\frac{\omega}{2} - (-1)^{(m+l-1)/2}i\sin^{m+l}\frac{\omega}{2}\right).\end{aligned}$$

Hence

$$|\Delta_{P_m,Q_{m,l}}(z)| = \left(\cos^{2(m+l)}\frac{\omega}{2} + \sin^{2(m+l)}\frac{\omega}{2}\right)^{1/2} \neq 0.$$

Therefore, we can define the following functions on the unit circle.

$$
\begin{aligned}
G_{m,l}(z) &= \frac{Q_{m,l}(-z)}{\Delta_{P_m,Q_{m,l}}(z)} = \begin{cases} \dfrac{2^m(1+z)^l}{(1+z)^{m+l}-(1-z)^{m+l}}, \\ \quad \text{when } m+l \text{ is odd, or } (m+l)/2 \text{ is odd}, \\ \dfrac{2^m(1+z)^l}{(1+z)^{m+l}+(1-z)^{m+l}}, \\ \quad \text{when } (m+l)/2 \text{ is even}, \end{cases} \\
H_{m,l}(z) &= \frac{-P_m(-z)}{\Delta_{P_m,Q_{m,l}}(z)} = \begin{cases} \dfrac{-2^l(1-z)^m}{(1+z)^{m+l}-(1-z)^{m+l}}, \\ \quad \text{when } m+l \text{ is odd, or } (m+l)/2 \text{ is odd}, \\ \dfrac{2^l z^{-1}(1-z)^m}{(1+z)^{m+l}+(1-z)^{m+l}}, \\ \quad \text{when } (m+l)/2 \text{ is even}. \end{cases}
\end{aligned} \tag{3.3}
$$

We define the following functions.

$$
\widehat{\widetilde{N}}_{m,l}(\omega) = \prod_{k=1}^{\infty} \overline{G_{m,l}(e^{-2^{-k}\omega i})}, \tag{3.4}
$$

$$
\widehat{\widetilde{\xi}}_{m,l}(\omega) = \overline{H_{m,l}(e^{-\frac{\omega i}{2}})}\widehat{\widetilde{N}}_{m,l}\left(\frac{\omega}{2}\right). \tag{3.5}
$$

By (3.3), we know that

$$
\overline{G_{m,l}(e^{-\omega i})} = \left(\frac{1+e^{-\omega i}}{2}\right)^l \widetilde{p}_{m,l}(\omega),
$$

where

$$
\widetilde{p}_{m,l}(\omega) = \begin{cases} \dfrac{2^{m+l}e^{-m\omega i}}{(1+e^{-\omega i})^{m+l}-(1-e^{-\omega i})^{m+l}}, & \text{when } (m+l)/2 \text{ is odd}, \\ \dfrac{2^{m+l}e^{-m\omega i}}{(1+e^{-\omega i})^{m+l}+(1-e^{-\omega i})^{m+l}}, & \text{when } m+l \text{ is odd, or } (m+l)/2 \text{ is even}. \end{cases} \tag{3.6}
$$

By (1.3) we know that the function $\widehat{N}_m$ satisfies the conditions in Theorem 1 with $q = 1$. We need to prove that the function $\widehat{\widetilde{N}}_{m,l}$ satisfies the conditions in Theorem 1 so that we can obtain the dual scaling function $\widetilde{N}_{m,l}$ and dual wavelet $\widetilde{\xi}_{m,l}$ via inverse Fourier transforms. We see that

$$
|\widetilde{p}_{m,l}(\omega)| = \begin{cases} \dfrac{1}{\sin^{m+l}(\omega/2)+\cos^{m+l}(\omega/2)}, & \text{when } m+l \text{ is even}, \\ \dfrac{1}{\left(\sin^{2(m+l)}(\omega/2)+\cos^{2(m+l)}(\omega/2)\right)^{1/2}}, & \text{when } m+l \text{ is odd}. \end{cases} \tag{3.7}
$$

THEOREM 2. *Let $\widetilde{p}_{m,l}$ be defined as in* (3.6). *If* $2 \le m+l \le 6$ *or* $m+l = 8, 10$,

then we have

$$|\widetilde{p}_{m,l}(\omega)| \le \left|\widetilde{p}_{m,l}\left(\frac{2\pi}{3}\right)\right|, \qquad 0 \le |\omega| \le \frac{\pi}{3}, \quad \frac{2\pi}{3} \le |\omega| \le \pi,$$

$$|\widetilde{p}_{m,l}(\omega)\widetilde{p}_{m,l}(2\omega)| \le \left|\widetilde{p}_{m,l}\left(\frac{2\pi}{3}\right)\right|^2, \qquad \frac{\pi}{3} \le |\omega| \le \frac{2\pi}{3}.$$

If $m+l = 7, 9$, *or* $m+l \ge 11$, *then we have* $[-\frac{2\pi}{3}, -\frac{\pi}{3}] \cup [\frac{\pi}{3}, \frac{2\pi}{3}] = D_1 \cup D_2$ *such that*

$$|\widetilde{p}_{m,l}(\omega)| \le \left|\widetilde{p}_{m,l}\left(\frac{2\pi}{3}\right)\right|, \qquad 0 \le |\omega| \le \frac{\pi}{3}, \quad \frac{2\pi}{3} \le |\omega| \le \pi,$$

$$|\widetilde{p}_{m,l}(\omega)\widetilde{p}_{m,l}(2\omega)| \le \left|\widetilde{p}_{m,l}\left(\frac{2\pi}{3}\right)\right|^2, \qquad \omega \in D_1,$$

$$|\widetilde{p}_{m,l}(\omega)\widetilde{p}_{m,l}(2\omega)\widetilde{p}_{m,l}(4\omega)| \le \left|\widetilde{p}_{m,l}\left(\frac{2\pi}{3}\right)\right|^3, \qquad \omega \in D_2.$$

We postpone the proof of Theorem 2 to the next section. It is easy to see that

$$\left|\widetilde{p}_{m,l}\left(\frac{2\pi}{3}\right)\right| = \begin{cases} \dfrac{2^{m+l}}{1+3^{(m+l)/2}} & \text{if } m+l \text{ is even,} \\ \dfrac{2^{m+l}}{\sqrt{1+3^{m+l}}} & \text{if } m+l \text{ is odd.} \end{cases} \tag{3.8}$$

Let

$$\alpha_{m,l} = l - \log_2 \left|\widetilde{p}_{m,l}\left(\frac{2\pi}{3}\right)\right| = \begin{cases} \log_2(1+3^{(m+l)/2}) - m & \text{if } m+l \text{ is even,} \\ \frac{1}{2}\log_2(1+3^{m+l}) - m & \text{if } m+l \text{ is odd.} \end{cases}$$

By Theorem 1 and Theorem 2 we have the next theorem.

THEOREM 3. *If* $\alpha_{m,l} > 1/2$, *then the functions* N_m *and* $\widetilde{N}_{m,l}$ *are dual scaling functions which generate the dual MRAs* $\{V_j^m\}$ *and* $\{\widetilde{V}_j^{m,l}\}$, *respectively. The functions* $\xi_{m,l}$ *and* $\widetilde{\xi}_{m,l}$ *are dual wavelets in the sense of Definition* 1. *In addition, if* $\alpha_{m,l} > 1$, *then the dual scaling function* $\widetilde{N}_{m,l}$ *and the dual wavelet* $\widetilde{\xi}_{m,l}$ *are in* $C^{\alpha_{m,l}-1}$.

Theorem 3 is mainly based on Theorem 1 and Theorem 2. It has been pointed out in [11] and [13] that the conditions in Theorem 1 are not strictly necessary to ensure that $\widehat{\phi}$ is in $L^2(\mathbf{R})$. But we can prove a weaker inverse theorem on the square integrability of the function $\widehat{\phi}$.

THEOREM 4. *Let*

$$m_0(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^L p(\omega), \qquad \widehat{\phi}(\omega) = \prod_{k=1}^{\infty} m_0(2^{-k}\omega).$$

Suppose that p *is continuous and*

$$|p(\omega)| = f\left(\sin^2\frac{\omega}{2}\right), \quad |p(\omega)| \ge 1, \quad \left|p\left(\frac{2\pi}{3}\right)\right| = 2^M, \quad M > L.$$

Then the function $\widehat{\phi}$ *is not in* $L^2(\mathbf{R})$.

Proof. Since p is continuous, we can choose ϵ between 0 and $M-L$ such that for some $\delta > 0$,

$$|p(\omega)| > 2^{L+\epsilon}, \qquad \omega \in \left[\frac{2\pi}{3} - \delta, \frac{2\pi}{3} + \delta\right].$$

For any nonnegative integer j we have

$$\begin{aligned}\sin\left(\frac{2\pi}{3}\cdot 2^j+\omega\right) &= \sin\left(\frac{2\pi}{3}\cdot 2^j\right)\cos\omega+\cos\left(\frac{2\pi}{3}\cdot 2^j\right)\sin\omega\\ &= (-1)^j\sin\frac{2\pi}{3}\cos\omega+\cos\frac{2\pi}{3}\sin\omega\\ &= (-1)^j\sin\left(\frac{2\pi}{3}+(-1)^j\omega\right).\end{aligned}$$

Hence,

$$|p(\omega)|>2^{L+\epsilon},\qquad \omega\in\left[\frac{2\pi}{3}\cdot 2^j-\delta,\frac{2\pi}{3}\cdot 2^j+\delta\right].$$

If $\omega\in[\frac{2\pi}{3}\cdot 2^n-\delta,\frac{2\pi}{3}\cdot 2^n+\delta]$, then $2^{-k}\omega\in[\frac{2\pi}{3}\cdot 2^{n-k}-2^{-k}\delta,\frac{2\pi}{3}\cdot 2^{n-k}+2^{-k}\delta]$, $k=1,2,\dots,n$. Consequently, we have

$$\begin{aligned}\int_{\frac{2\pi}{3}\cdot 2^n-\delta}^{\frac{2\pi}{3}\cdot 2^n+\delta}|\widehat{\phi}(\omega)|^2\,d\omega &= \int_{\frac{2\pi}{3}\cdot 2^n-\delta}^{\frac{2\pi}{3}\cdot 2^n+\delta}\left|\frac{\sin(\omega/2)}{\omega/2}\right|^{2L}\prod_{k=1}^{\infty}|p(2^{-k}\omega)|^2\,d\omega\\ &\geq \int_{\frac{2\pi}{3}\cdot 2^n-\delta}^{\frac{2\pi}{3}\cdot 2^n+\delta}\left|\frac{\sin(\omega/2)}{\omega/2}\right|^{2L}\prod_{k=1}^{n}|p(2^{-k}\omega)|^2\,d\omega\\ &\geq C_1\int_{\frac{2\pi}{3}\cdot 2^n-\delta}^{\frac{2\pi}{3}\cdot 2^n+\delta}\left(\frac{1}{\frac{2\pi}{3}\cdot 2^n+\delta}\right)^{2L}(2^{L+\epsilon})^{2n}\,d\omega\\ &= 2C_1\delta\left(\frac{2^n}{\frac{2\pi}{3}\cdot 2^n+\delta}\right)^{2L}\cdot 2^{2n\epsilon}>C_2,\end{aligned}$$

where C_1 and C_2 are positive constants that do not depend on n. The last inequality is true for sufficiently large n. Hence,

$$\int_{-\infty}^{\infty}|\widehat{\phi}(\omega)|^2\,d\omega\geq\sum_{n=1}^{\infty}\int_{\frac{2\pi}{3}\cdot 2^n-\delta}^{\frac{2\pi}{3}\cdot 2^n+\delta}|\widehat{\phi}(\omega)|^2\,d\omega=+\infty.$$

Since the function $\widehat{\widetilde{N}}_{m,l}$ satisfies all the conditions in Theorem 4 and $|\widetilde{p}_{m,l}(\frac{2\pi}{3})|=2^{l-\alpha_{m,l}}$, we derive the following theorem.

THEOREM 5. *If $\alpha_{m,l}<0$, then the functions $\widehat{\widetilde{N}}_{m,l}$ and $\widehat{\widetilde{\xi}}_{m,l}$ are not in $L^2(\mathbf{R})$. Consequently, the function $\xi_{m,l}$ is not a wavelet when $\alpha_{m,l}<0$.*

There is still a gap: "$0\leq\alpha_{m,l}\leq 1/2$." However, this gap is not too large. It is easy to see that for any fixed m, there is at most one integer l such that $\alpha_{m,l}$ is in this gap. When $1\leq m\leq 20$, one can verify that only the following twelve $\alpha_{m,l}$'s are in the interval $[0,1/2]$.

$$\begin{array}{llll}\alpha_{2,1}=0.403677, & \alpha_{3,1}=0.321928, & \alpha_{6,2}=0.357552, & \alpha_{7,2}=0.132368,\\ \alpha_{10,3}=0.302257, & \alpha_{11,3}=0.095397, & \alpha_{13,4}=0.472181, & \alpha_{14,4}=0.264736,\\ \alpha_{15,4}=0.0571438, & \alpha_{17,5}=0.434596, & \alpha_{18,5}=0.227069, & \alpha_{19,5}=0.0195527.\end{array}$$

We have seen that the length of the support of the wavelet $\xi_{m,l}$ is $(m+l)/2$. When $l<m$, the length of the support of the wavelet $\xi_{m,l}$ is even less than the length of the support of the scaling function N_m. The smoothness of the dual wavelet $\widetilde{\xi}_{m,l}$ is $\alpha_{m,l}-1$. The smaller l is, the smaller the support of the wavelet $\xi_{m,l}$. The larger l is, the better the smoothness of the dual wavelet $\widetilde{\xi}_{m,l}$. For applications, if one wants

the small support of the wavelet, one can choose small l. If one wants the better smoothness of the dual wavelet, one can choose large l.

Although the dual scaling functions and dual wavelets are noncompactly supported, the algorithms should be still manageable since the dual filters are fractions of trigonometric polynomials and can be implemented in a fast recursive way.

We compare the wavelet $\xi_{m,l}$ with the wavelet $\psi_{m,l}$. In [3] we proved that the function

$$\psi_{m,l}(x) = \frac{1}{2^{l-1}} \sum_{j=0}^{m+l-2} (-1)^j N_{m+l}(j+1) N_{m+l}^{(l)}(2x-j)$$

is also an mth-order spline wavelet. The support of the wavelet $\psi_{m,l}$ is $[0, m+l-1]$. The dual wavelet $\widetilde{\psi}_{m,l}$ is an lth-order spline function. If $l_1 = m + 2l_2 - 2$, then the functions ξ_{m,l_1} and ψ_{m,l_2} have the same length of the support. The smoothness of the wavelet ξ_{m,l_1} is

$$\begin{aligned}\alpha_{m,l_1} - 1 &= \log_2(1 + 3^{(m+m+2l_2-2)/2}) - m - 1 = \log_2(1 + 3^{(m+l_2-1)}) - m - 1 \\ &> (m + l_2 - 1)\log_2 3 - m - 1 = l_2 \log_2 3 + (m-1)(\log_2 3 - 1) - 2,\end{aligned}$$

which is much better than the smoothness of the wavelet ψ_{m,l_2}, except when $m = l_1 = l_2 = 1$. In this aspect, the wavelets $\xi_{m,l}$ are better than the wavelets $\psi_{m,l}$. But the dual wavelets $\widetilde{\xi}_{m,l}$ are no longer spline functions except when $m = l = 1$. In this aspect, the wavelets $\psi_{m,l}$ are better than the wavelets $\xi_{m,l}$.

At the end of this section we mention an interesting fact. Obviously, for any integer j, the function $\psi(\cdot - j)$ is a wavelet if the function ψ is a wavelet. In general, the function $\psi(\cdot - 1/2)$ is not a wavelet. For example, when $m+l$ is even and $\alpha_{m,l} > 1/2$, the function $\xi_{m,l}$ is a wavelet but the function $\xi_{m,l}(\cdot - 1/2)$ is not a wavelet. But when $m + l$ is odd and $\alpha_{m,l} > 1/2$, both $\xi_{m,l}$ and $\zeta_{m,l}(x) = \xi_{m,l}(x - 1/2)$ are wavelets. In fact, we have

$$\widehat{\zeta}_{m,l}(\omega) = Q_{m,l}^*(e^{-\omega i/2})\widehat{N}_m\left(\frac{\omega}{2}\right) = 2^{-l}e^{-\omega i/2}(1 - e^{-\omega i/2})^l \widehat{N}_m\left(\frac{\omega}{2}\right)$$

and

$$|\Delta_{P_m, Q_{m,l}^*}(z)| = |\Delta_{P_m, Q_{m,l}}(z)| \neq 0, \qquad |z| = 1.$$

Hence, we can prove that $\zeta_{m,l}(x) = \xi_{m,l}(x - 1/2)$ is a wavelet when $\alpha_{m,l} > 1/2$.

When $m + l$ is even, the center of the wavelet $\xi_{m,l}$ is an integer or half-integer. When $m + l$ is odd, the center of the wavelet $\xi_{m,l}$ is the fourth integer $(m + l)/4$.

4. The proof of Theorem 2. We have seen that Theorem 2 is essential in our constructions. In this section we shall prove Theorem 2. Recall that

$$|\widetilde{p}_{m,l}(\omega)| = \begin{cases} \dfrac{1}{\sin^{m+l}(\omega/2) + \cos^{m+l}(\omega/2)}, & \text{when } m+l \text{ is even,} \\ \dfrac{1}{\left(\sin^{2(m+l)}(\omega/2) + \cos^{2(m+l)}(\omega/2)\right)^{1/2}}, & \text{when } m+l \text{ is odd.} \end{cases} \tag{4.1}$$

Let

$$\sin^2(\omega/2) = x, \quad f(x) = 4x(1-x), \quad f_2(x) = f(f(x)),$$

$$W_p(x) = \frac{1}{x^p + (1-x)^p}, \quad F_p(x) = \frac{1}{W_p(x)} = x^p + (1-x)^p.$$

If $m+l$ is even, by setting $p=(m+l)/2$ we have

$$|\widetilde{p}_{m,l}(\omega)| = W_p(x), \quad |\widetilde{p}_{m,l}(2\omega)| = W_p(f(x)), \quad |\widetilde{p}_{m,l}(4\omega)| = W_p(f_2(x)).$$

If $m+l$ is odd, by setting $p=m+l$ we have

$$|\widetilde{p}_{m,l}(\omega)| = \sqrt{W_p(x)}, \quad |\widetilde{p}_{m,l}(2\omega)| = \sqrt{W_p(f(x))}, \quad |\widetilde{p}_{m,l}(4\omega)| = \sqrt{W_p(f_2(x))}.$$

It is easy to see that Theorem 2 is equivalent to the following theorems.

THEOREM 6. *Let $1 \le p \le 5$. We have*

$$F_p(x) \ge F_p\left(\frac{3}{4}\right), \quad \textit{when} \quad 0 \le x \le \frac{1}{4}, \quad \frac{3}{4} \le x \le 1, \tag{4.2}$$

$$F_p(x)F_p(f(x)) \ge F_p^2\left(\frac{3}{4}\right), \quad \textit{when} \quad \frac{1}{4} \le x \le \frac{3}{4}. \tag{4.3}$$

THEOREM 7. *Let $p \ge 6$. We have*

$$F_p(x) \ge F_p\left(\frac{3}{4}\right), \quad \textit{when} \quad 0 \le x \le \frac{1}{4}, \quad \frac{3}{4} \le x \le 1, \tag{4.4}$$

$$F_p(x)F_p(f(x)) \ge F_p^2\left(\frac{3}{4}\right), \quad \textit{when} \quad \frac{1}{4} \le x \le 0.41, \quad 0.59 \le x \le \frac{3}{4}, \tag{4.5}$$

$$F_p(x)F_p(f(x))F_p(f_2(x)) \ge F_p^3\left(\frac{3}{4}\right), \quad \textit{when} \quad 0.41 \le x \le 0.59. \tag{4.6}$$

Since $F_p(x) = F_p(1-x)$ and $f(x) = f(1-x)$, we only need to prove the above theorems for $0.5 \le x \le 1$. First we prove Theorem 6.

Proof of Theorem 6. Since the function F_p is increasing on $[1/2, 1]$, the inequality (4.2) is obvious.

Let

$$A_p(x) = F_p(x)F_p(f(x)) = (x^p + (1-x)^p)(f^p(x) + (1-f(x))^p).$$

The derivative of the function A_p is

$$\begin{aligned} A_p'(x) =& p(x^{p-1} - (1-x)^{p-1})(f^p(x) + (1-f(x))^p) \\ & - 4p(x^p + (1-x)^p)(f^{p-1}(x) - (1-f(x))^{p-1})(2x-1). \end{aligned} \tag{4.7}$$

When $p=1$, the function F_1 is the constant 1, and the inequality (4.3) is obvious. When $p=2,3$, we have

$$\begin{aligned} A_2'(x) &= 2(2x-1)(6(1-f(x))^2 - 1), \\ A_3'(x) &= 3(2x-1)^3(6x-1)(6x-5). \end{aligned}$$

When $1/2 \le x \le 3/4$, we see that $3/4 \le f(x) \le 1$. One can verify that the functions A_2' and A_3' are negative on $(1/2, 3/4]$. Hence, the functions A_2 and A_3 are decreasing on $[1/2, 3/4]$, which proves inequality (4.3) for $p=2,3$.

Since the functions $(x^{p-1} - (1-x)^{p-1})/(2x-1)$ and $x^p + (1-x)^p$ can be expressed as polynomials in $f(x)$, the function $A_p'(x)/(2x-1)$ can be expressed as a polynomial

in $f(x)$ also. By a simple calculation we have

$$\begin{aligned} A_4'(x) &= -(2x-1)(6f^5(x) - 50f^4(x) + 108f^3(x) - 126f^2(x) + 81f(x) - 20) \\ &= -(2x-1)B_4(f(x)) \\ A_5'(x) &= -\frac{5}{4}(2x-1)(30f^5(x) - 150f^4(x) + 264f^3(x) + 231f^2(x) + 106f(x) - 20) \\ &= -\frac{5}{4}(2x-1)B_5(f(x)). \end{aligned}$$

The zeros of the polynomials B_4 and B_5 are

$$0.537296, \quad 0.542624 \pm 0.927367i, \quad 0.92954, \quad 5.78125$$

and

$$0.52765, \quad 0.528725 \pm 0.525581i, \quad 0.906141, \quad 2.50876.$$

Looking at these zeros, we see that the polynomial B_p, $p = 4, 5$ has only one zero in $[3/4, 1]$. Because the function f is decreasing on $[1/2, 1]$ and $f(1/2) = 1$, and $f(3/4) = 3/4$, the function A_p', $p = 4, 5$ has only one zero x_p in $(1/2, 3/4]$. Since

$$\begin{aligned} A_p'(3/4) &= -p((3/4)^{p-1} - (1/4)^{p-1})((3/4)^p + (1/4)^p) < 0, \\ \lim_{x \to 1/2} \frac{A_p'(x)}{2x-1} &= p(p-1)(1/2)^{p-2} - p(1/2)^{p-3} > 0, \end{aligned}$$

the function A_p, $p = 4, 5$ is increasing on $[1/2, x_p]$ and decreasing on $[x_p, 3/4]$. So when $p = 4, 5$ and $1/2 \le x \le 3/4$, we have

$$\begin{aligned} A_p(x) &\ge \min[A_p(1/2), A_p(3/4)] \\ &= \min[(1/2)^{p-1}, ((3/4)^p + (1/4)^p)^2] = ((3/4)^p + (1/4)^p)^2 = F_p^2(3/4) \end{aligned}$$

which proves inequality (4.3) for $p = 4, 5$. □

When $p \ge 6$, one can verify that $A_p(1/2) < A_p(3/4)$. That is why we need to consider the function $F_p(x)F_p(f(x))F_p(f_2(x))$ when $p \ge 6$. Before we prove Theorem 7, we establish the following lemma.

LEMMA 2. *The function $A_p(x) = F_p(x)F_p(f(x))$ is decreasing on $[0.7, 3/4]$.*

Proof. In the proof of Theorem 6, we have already seen that this conclusion is true for $p = 1, 2, 3$. In the following we shall prove that this conclusion is true for $p \ge 4$. In fact, in Theorem 7, we only need this lemma for $p \ge 6$. Recall that

$$\begin{aligned} (1/p)A_p'(x) = &(x^{p-1} - (1-x)^{p-1})(f^p(x) + (1-f(x))^p) \\ &- 4(x^p + (1-x)^p)(f^{p-1}(x) - (1-f(x))^{p-1})(2x-1). \end{aligned}$$

When $0.7 \le x \le 3/4$, we have $2x - 1 \ge 0.4$. Since the functions $x^p + (1-x)^p$ and $f^{p-1}(x) - (1-f(x))^{p-1}$ are positive on $[1/2, 3/4]$, we have

(4.8)

$$\begin{aligned} &(1/p)A_p'(x) \\ &\le (x^{p-1} - (1-x)^{p-1})(f^p(x) + (1-f(x))^p) - 1.6(x^p + (1-x)^p)(f^{p-1}(x) - (1-f(x))^{p-1}). \end{aligned}$$

When $0.7 \le x \le 3/4$, we have $3/4 \le f(x) \le f(0.7) = 0.84$. Hence,

$$\begin{aligned} & f^p(x) + (1-f(x))^p - (f^{p-1}(x) - (1-f(x))^{p-1}) \\ & = f^{p-1}(f(x)-1) + (1-f(x))^p + (1-f(x))^{p-1} \\ & \le -0.16 f^{p-1}(x) + (1/4)^p + (1/4)^{p-1} \\ & \le -0.16(3/4)^{p-1} + (5/4)(1/4)^{p-1} \\ & = (-0.16 \cdot 3^{p-1} + 5/4)(1/4)^{p-1} < 0. \end{aligned} \tag{4.9}$$

The last inequality is true for $p \ge 3$. Also we have

$$x^{p-1} - (1-x)^{p-1} - 1.6(x^p + (1-x)^p) < x^{p-1} - 1.6x^p = x^{p-1}(1-1.6x) < -0.12x^{p-1} < 0. \tag{4.10}$$

Combining the inequalities (4.8)–(4.10), we see that $A_p'(x) < 0$ when $0.7 \le x \le 3/4$. Hence, the function A_p is decreasing on $0.7 \le x \le 3/4$. □

Proof of Theorem 7. Since the function F_p is increasing on $[1/2, 1]$, the inequality (4.4) is obvious.

By Lemma 2, the function $F_p(x)F_p(f(x))$ is decreasing on $[0.7, 3/4]$. This proves the inequality (4.5) for $0.7 \le x \le 3/4$.

For $0.59 \le x \le 0.7$ we use a different strategy. When $p \ge 6$, we have

$$(3.001)^p = 3^p + p \cdot 3^{p-1} \cdot 0.001 + \cdots > 3^p + 6 \cdot 3^5 \cdot 0.001 > 3^p + 1. \tag{4.11}$$

Then we have

$$F_p^2(3/4) = ((3/4)^p + (1/4)^p)^2 < \left(\frac{3.001}{4}\right)^{2p} = \left(\frac{9.006001}{16}\right)^p < 0.5629^p. \tag{4.12}$$

It is easy to see that the function F_p is increasing on $[1/2, 1]$ and the function $F_p(f(x))$ is decreasing on $[1/2, (2+\sqrt{2})/4]$. When $x \in [a, b] \subset [1/2, (2+\sqrt{2})/4]$ we have

$$F_p(x)F_p(f(x)) \ge F_p(a)F_p(f(b)) = (a^p + (1-a)^p)(f^p(b) + (1-f(b))^p) > (af(b))^p. \tag{4.13}$$

By using the above fact, we can prove that the inequality (4.5) is true for $0.59 \le x \le 0.7$. In fact, when $0.59 \le x \le 0.6$ we have

$$F_p(x)F_p(f(x)) > (0.59 \cdot f(0.6))^p = (0.59 \cdot 0.96)^p = 0.5664^p > 0.5629^p > F_p^2(3/4).$$

When $0.6 \le x \le 0.62$ we have

$$F_p(x)F_p(f(x)) > (0.6 \cdot f(0.62))^p = (0.6 \cdot 0.9424)^p = 0.56544^p > 0.5629^p > F_p^2(3/4).$$

When $0.62 \le x \le 0.65$ we have

$$F_p(x)F_p(f(x)) > (0.62 \cdot f(0.65))^p = (0.62 \cdot 0.91)^p = 0.5642^p > 0.5629^p > F_p^2(3/4).$$

When $0.65 \le x \le 0.68$ we have

$$F_p(x)F_p(f(x)) > (0.65 \cdot f(0.68))^p = (0.65 \cdot 0.8704)^p = 0.56576^p > 0.5629^p > F_p^2(3/4).$$

When $0.68 \le x \le 0.7$ we have

$$F_p(x)F_p(f(x)) > (0.68 \cdot f(0.7))^p = (0.68 \cdot 0.84)^p = 0.5712^p > 0.5629^p > F_p^2(3/4).$$

In the rest of the proof, we need to prove the inequality (4.6). First, by inequality (4.11), we have

$$F_p^3(3/4) = \left(\frac{3^p+1}{4^p}\right)^3 < \left(\frac{3.001}{4}\right)^{3p} < 0.4223^p.$$

Note these facts: The function F_p is increasing on $[1/2, 1]$. The function $F_p(f(x))$ is decreasing on $[1/2, (2+\sqrt{2})/4]$. The function $F_p(f_2(x))$ is decreasing on $[1/2, (2+\sqrt{2-\sqrt{2}})/4]$. When $1/2 \le x \le 0.59$, we have

$$\begin{aligned} F_p(x)F_p(f(x))F_p(f_2(x)) &\ge F_p(1/2)F_p(f(0.59))F_p(f_2(0.59)) \\ &> (1/2)^p \cdot f^p(0.59) \cdot (1 - f_2(0.59))^p \\ &= (0.5 \cdot 0.9676 \cdot 0.87459904)^p = (0.423131015552)^p \\ &> 0.4223^p > F_p^3(3/4). \end{aligned}$$

Finally, we must mention that all the above numbers are accurate. We did not perform any rounding. □

5. Final remarks. In [3] and in this paper we proved that the functions $\psi_{m,l;c}$, $L^{(l)}_{m+l;c}(2x-1)$, and $\xi_{m,l}$ are mth-order spline wavelets. In fact, every mth-order cardinal spline wavelet can be written as

$$\psi_a(x) = \frac{1}{2^{l-1}} \sum_j a_j N^{(l)}_{m+l}(2x-j), \qquad \sum_j a_j \neq 0,$$

where $a = \{a_j\}$ is a real sequence, whether finite or infinite. The function ψ_a is also dependent on m and l, but we omit the subscripts m and l. The problem is how to choose these a_j's to ensure that the function ψ_a is a wavelet. First, we need to choose $a = \{a_j\}$ to be in l^1. The Fourier transform of ψ_a is

$$\widehat{\psi}_a(\omega) = Q_a(e^{-\omega i/2})\widehat{N}_m\left(\frac{\omega}{2}\right),$$

where

$$Q_a(z) = \left(\frac{1-z}{2}\right)^l \sum_j a_j z^j.$$

Second, we must choose $\{a_j\}$ such that the determinant

$$\Delta_{P_m,Q_a}(z) = \begin{vmatrix} P_m(z) & Q_a(z) \\ P_m(-z) & Q_a(-z) \end{vmatrix}$$

is not equal to zero on the unit circle. Then we can define

$$G_a(z) = \frac{Q_a(-z)}{\Delta_{P_m,Q_a}(z)}, \qquad H_a(z) = \frac{-P_m(-z)}{\Delta_{P_m,Q_a}(z)},$$

$$\widehat{\widehat{N}}_a(\omega) = \prod_{k=1}^{\infty} \overline{G_a(e^{-2^{-k}\omega i})}, \qquad \widehat{\widehat{\psi}}_a(\omega) = \overline{H_a(e^{-\frac{\omega i}{2}})}\widehat{\widehat{N}}_a\left(\frac{\omega}{2}\right).$$

Third, we need to prove that the function $\overline{G_a}$ satisfies the hypotheses of Cohen, Daubechies, and Feauveau's theorems so that we can obtain the dual scaling function and dual wavelet via inverse Fourier transforms. To construct the wavelets $\xi_{m,l}$ and

$\psi_{m,l;c}$, we used different strategies to prove that the corresponding functions $\overline{G_a}$'s satisfy the hypotheses of Cohen, Daubechies, and Feauveau's theorems. If the sequence $\{a_j\}$ is symmetric, then the wavelet and dual wavelet are symmetric (if l is even) or antisymmetric (if l is odd). If the sequence $\{a_j\}$ is finite and if the determinant $\Delta_{P_m,Q_a}(z)$ is a monomial, then both wavelet and dual wavelet are compactly supported. The authors of [13] consider mainly the case in which $m + l$ is even. We also can construct compactly supported wavelets when $m + l$ is odd. At least, it is interesting from a theoretical viewpoint. If

$$\widehat{\widetilde{N}}_a(\omega) = \prod_{k=1}^{\infty} \overline{G_a(e^{-2^{-k}\omega i})} = A_a(e^{-\omega i})\widehat{N}_l(\omega),$$

where A_a is a Laurent series, then the dual wavelet is an lth-order spline function [3]. In this case both wavelet and dual wavelet are spline functions, and we can obtain biorthogonal spline wavelet basis. Furthermore, if $m = l$, then the wavelet spaces W_j^m are kept orthogonal and the wavelet basis is nonorthogonal only inside a given scale.

Finally, we mention that our method also can be used to construct nonspline wavelets.

Appendix. In illustration of our techniques, we present some graphs of the scaling functions, the wavelets, the dual scaling functions, and the dual wavelets for $m = 1$ (Fig. 1), $m = 2$ (Fig. 2), and $m = 3$ (Fig. 3).

Acknowledgments. I am deeply grateful to Ward Cheney, my supervisor, for his encouragement, patience, and assistance throughout my years in Austin, and for his guidance in my research and writing of this paper. From the very begining of my research for my Ph.D. dissertation, he has strongly encouraged and advised me to study wavelet theory. I wish to express special thanks to John Gilbert for his many suggestions and help in my work. I appreciate the opportunity as a teaching assistant for his class on wavelets, which helped me in both teaching and research. I am also indebted to Ingrid Daubechies,Charles Chui,and Jian-Zhong Wang for their advice, help, and suggestions. In particular, I wish to acknowledge the generosity of Jian-Zhong Wang for sharing with me his ideas and knowledge about wavelet theory. Finally, I thank Albert Cohen and the referees for their comments and suggestions that improved the presentation of this paper.

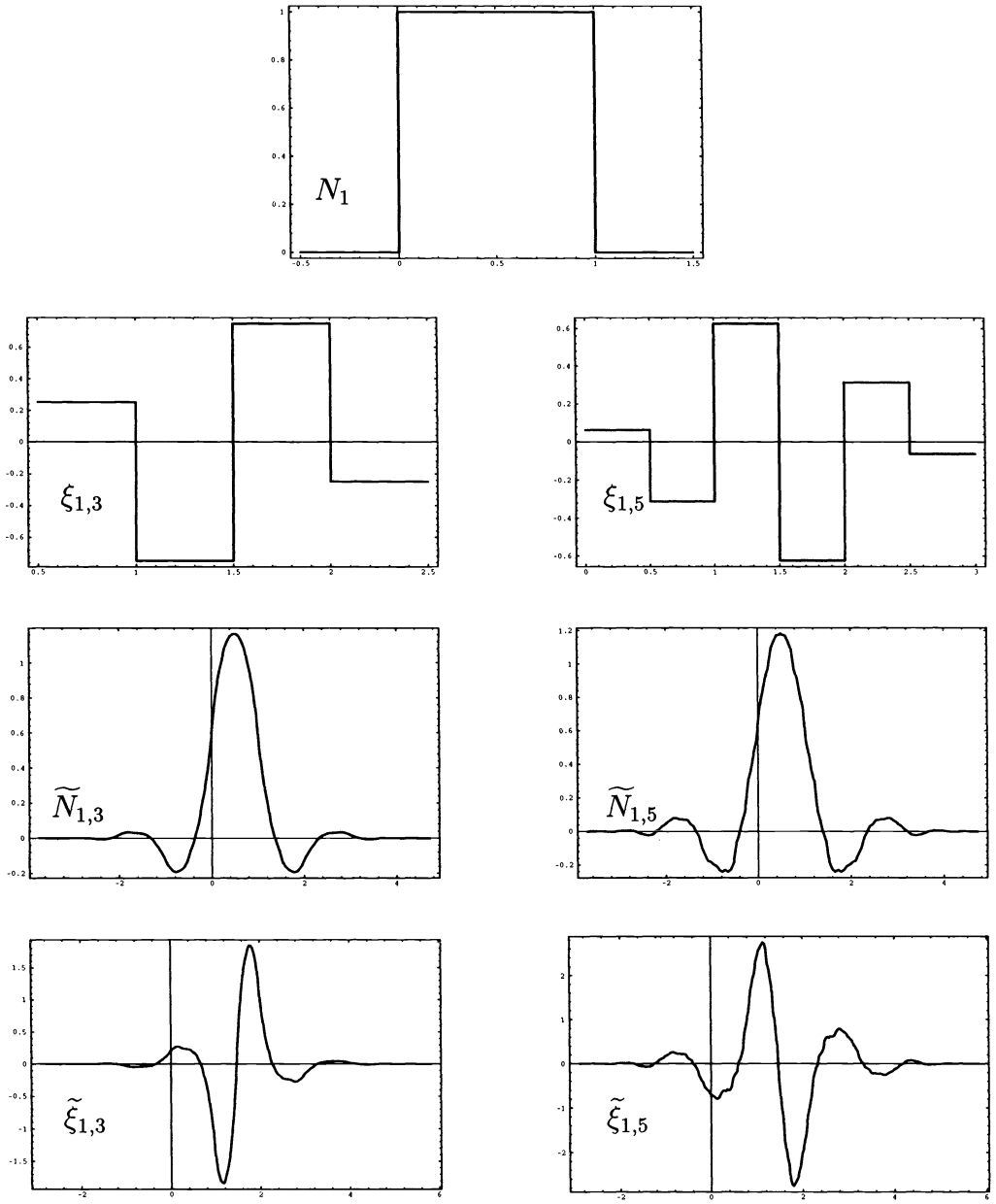


FIG. 1. *Graphs of the scaling function, the wavelets, the dual scaling functions, and the dual wavelets for* $m = 1$.

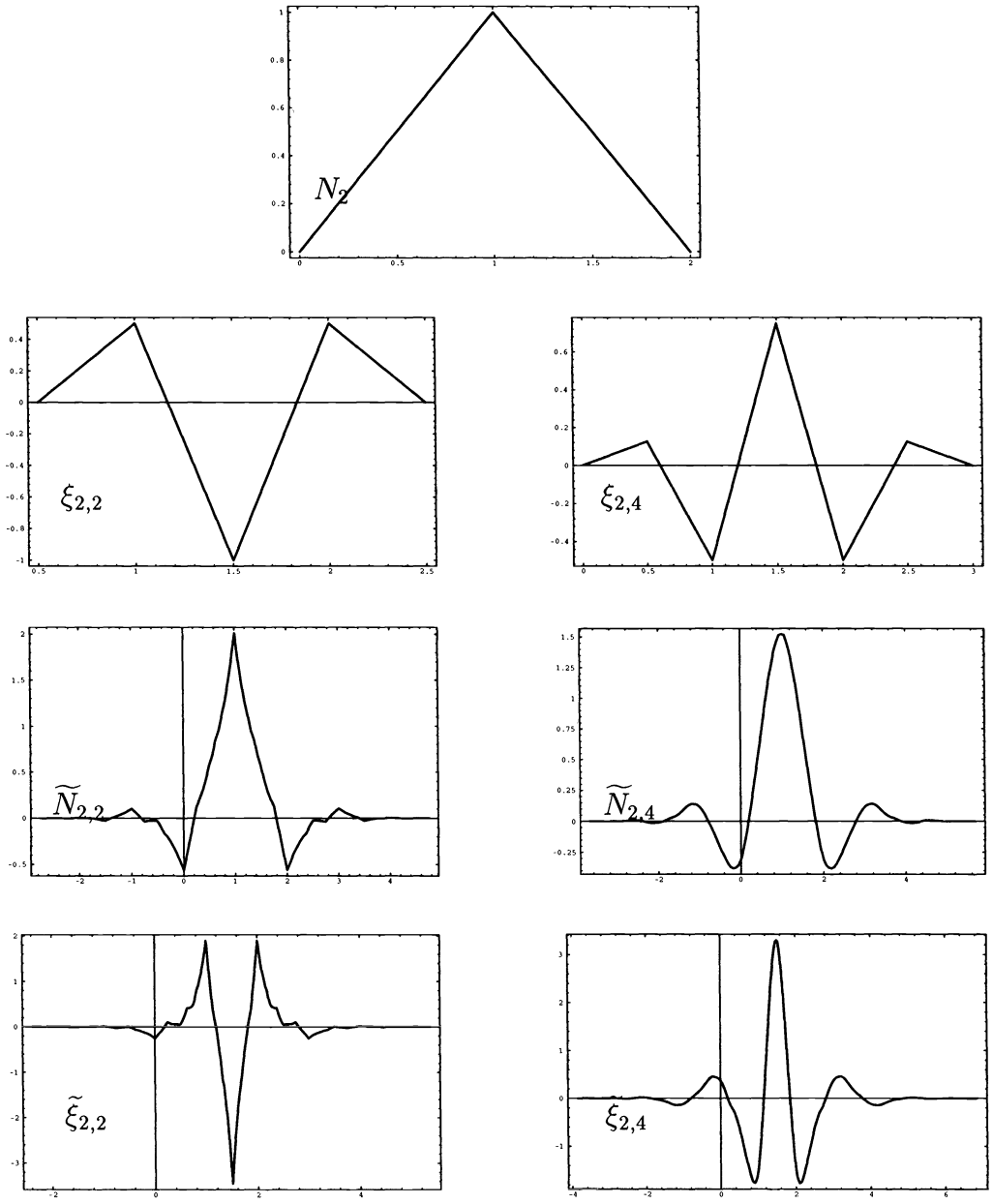


FIG. 2. *Graphs of the scaling function, the wavelets, the dual scaling functions, and the dual wavelets for* $m = 2$.

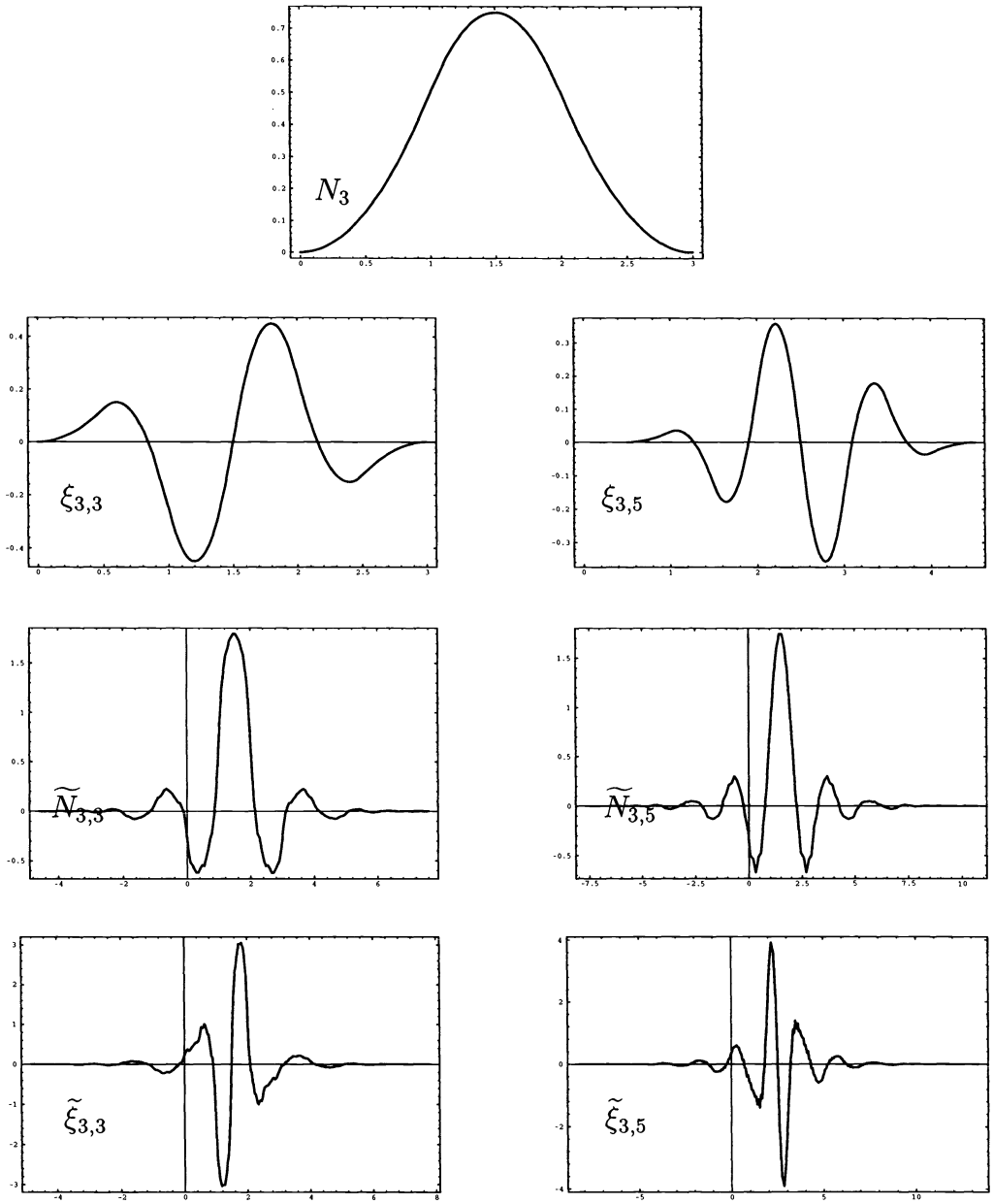


FIG. 3. *Graphs of the scaling function, the wavelets, the dual scaling functions, and the dual wavelets for* $m = 3$.

REFERENCES

[1] P. AUSCHER, *Ondelettes Fractales et Applications*, Ph.D. Thesis, Univ. Paris-Dauphine, 1989.

[2] G. BATTLE, *A block spin construction of ondelettes. Part I: Lemarie functions*, Comm. Math. Phys., 110 (1987), pp. 601-615.

[3] D. CHEN, *Extended families of cardinal spline wavelets*, Appl. Comp. Harmonic Anal., 1 (1994), pp. 194-208.

[4] C. K. CHUI, *An Introduction to Wavelets*, Academic Press, Boston, 1992.

[5] C. K. CHUI AND X. L. SHI, *Inequalities of Littlewood-Paley type for frames and wavelets*, CAT Report 249, Texas A&M Univ., College Station, TX, 1991.

[6] ———, *On a Littlewood-Paley identity and characterization of wavelets*, CAT Report 250, Texas A&M Univ., College Station, TX, 1991.

[7] C. K. CHUI AND J. Z. WANG, *A cardinal spline approach to wavelets*, Proc. Amer. Math. Soc., 113 (1991), pp. 785-793.

[8] ———, *On compactly supported spline wavelets and a duality principle*, Trans Amer. Math. Soc., 330 (1992), pp. 903-915.

[9] ———, *A general framework of compactly supported splines and wavelets*, J. Approx. Theory, 71 (1992), pp. 263–304.

[10] A. COHEN, *Ondelettes, Analyses Multiresolutions et Traitement Numerique du Signal*, Ph.D. Thesis, Univ. Paris- Dauphine, 1990.

[11] ———, *Biorthogonal wavelets*, in Wavelets—A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, Boston, 1992, pp. 123-152.

[12] A. COHEN AND I. DAUBECHIES, *Nonseparable bidimensional wavelet bases*, AT&T Bell Laboratories, preprint, 1991.

[13] A. COHEN, I. DAUBECHIES, AND J. C. FEAUVEAU, *Bi-orthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485-560.

[14] A. COHEN AND J. M. SCHLENKER, *Compactly supported bidimensional wavelet bases with hexagonal symmetry*, preprint, 1991, AT&T Bell Laboratories.

[15] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909-996.

[16] ———, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Trans. Inform. Theory, 36 (1990), pp. 961-1005.

[17] ———, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. Appl. Math. 61, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[18] J. C. FEAUVEAU, *Nonorthogonal Analysis Using Wavelets*, in Wavelets—A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, Boston, 1992, pp. 153-178.

[19] J. GILBERT, *Wavelets: Theory and Applications*, lecture notes, Mathematics Department, University of Texas at Austin , 1991.

[20] P. G. LEMARIÉ, *Une nouvelle base d'ondelettes de* $L^2(\mathbf{R}^n)$, J. Math. Pures Appl., 67 (1988), pp. 227-236.

[21] S. MALLAT, *Multiresolution approximations and orthonormal bases of* $L^2(\mathbf{R})$, Trans. Amer. Math. Soc., 315 (1989), pp. 69-87.

[22] S. MALLAT AND S. ZHONG, *Characterization of signals from multiscale edges*, IEEE Trans. PAMI, 14 (1992), pp. 710-732.

[23] Y. MEYER, *Ondelettes et Opérateurs. vol.*1, Hermann, Paris, 1990. (In French.)

[24] C. A. MICCHELLI, *Using the refinement equation for the construction of pre-wavelets*, Numer. Algorithms, 1 (1991), pp. 75-116.

[25] I. J. SCHOENBERG, *Cardinal Spline Interpolation*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 12, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1973.

[26] L. L. SCHUMAKER, *Spline Functions: Basic Theory*, John Wiley and Sons, New York, 1981.

[27] M. UNSER AND A. ALDROUBI, *Polynomial splines and wavelets—A signal processing Perspective*, in Wavelets—A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, Boston, 1992, pp. 91-121.

[28] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 519–528, May 1995

NONLINEAR STABILITY OF STRONG DETONATIONS FOR A VISCOUS COMBUSTION MODEL *

TAI-PING LIU† AND LONG-AN YING‡

Abstract. Strong detonations for a viscous combustion model are studied. These waves are of ZND (Zeldovich–von Neumann–Doring) type. It is shown that these waves are nonlinearly stable. The analysis consists of an energy method for the fluid variable and a pointwise estimate for the reactant. Strong detonations are compressive, which allows for a priori determination of their time-asymptotic location by the conservation law for the perturbation.

Key words. detonation waves, nonlinear stability

AMS subject classifications. 35K55, 76N10

1. Introduction.

Consider the viscous combustion model:

$$(u+qz)_t + f(u)_x = \beta u_{xx}, \tag{1.1}_1$$

$$z_t = -K\phi(u)z \tag{1.1}_2$$

[1], [3]. Here u represents the lumped fluid variables and z is the concentration of the reactant. The viscosity β, the heat release q, and the reaction rate K are positive constants. Motivated by the study of shock waves for gas dynamics, we require the flux $f(u)$ to satisfy

$$f'(u) > 0, \qquad f''(u) > 0 \tag{1.2}$$

so that (1.1) models detonations instead of deflagrations. The concentration z lies between 0 and 1, with $z = 1$ the unburnt state and $z = 0$ the completely burnt state. The reaction rate function $\phi(u)$ is smooth and satisfies

$$\phi(u) = \begin{cases} 1 & \text{for } u > 0, \\ 0 & \text{for } u < 0, \end{cases}$$
$$\phi'(u) > 0 \quad \text{for } 0 < u < 1, \tag{1.3}$$

so that $u = 0$ is the ignition temperature.

Well-posedness of the initial-value problem for (1.1) has been established [4], [5]. Our purpose is to study the nonlinear stability of strong detonation waves. There are two types of combustion waves for (1.1) and (1.2), and they are strong and weak detonation waves. Both are travelling waves:

$$(u, qz)(x,t) = (s,y)(x-\sigma t) \equiv (s, q\zeta)(x-\sigma t),$$
$$s(-\infty) = u_l, \qquad s(+\infty) = u_r,$$
$$\zeta(-\infty) = 0, \qquad \zeta(+\infty) = 1. \tag{1.4}$$

* Received by the editors November 15, 1993; accepted for publication November 15, 1993.

† Department of Mathematics, Stanford University, California 94305. This research was supported in part by the Army Basic Research grant and a National Science Foundation grant.

‡ Department of Mathematics, Peking University, Beijing, China 100871. This research was supported in part by the National Science Foundation grant through the U.S.–China Program.

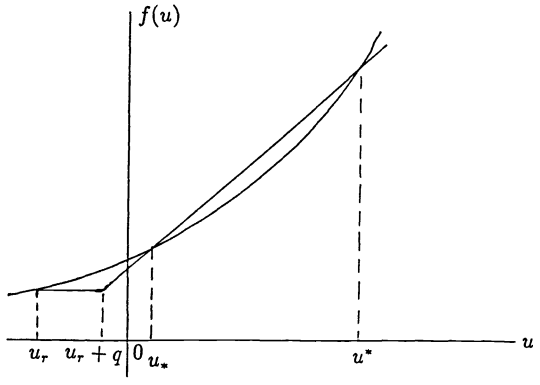


FIG. 1. *Chapman–Jouget diagram for detonations.*

By putting (1.4) into (1.1), we obtain ordinary differential equations for (s, y). The first equation in (1.1) is a conservation law and can be integrated to yield the jump condition

$$\sigma = \frac{f(u_r) - f(u_l)}{u_r + q - u_l}. \tag{1.5}$$

In general there are two values, u_* and u^*, for the left state u_l for each given right state u_r and wave speed σ; see Fig. 1.

The main difference between the weak detonation (u_*, u_r) and the strong detonation (u^*, u_r) is that the former is undercompressive and the latter is not:

$$\sigma > f'(u_*),$$

$$f'(u^*) > \sigma > f'(u_r). \tag{1.6}$$

The strong compressibility property (1.6) allows us to determine a priori the time-asymptotic location of a strong detonation when perturbed. This is done using the conservation laws

$$\int_{-\infty}^{\infty} (u + qz - (s + y))(x, t)\, dx = \int_{-\infty}^{\infty} (u + qz - (s + y))(x, 0)\, dx, \qquad t \geq 0. \tag{1.7}$$

The reason this can be done is that no characteristic leaves the detonation and therefore a perturbation does not give rise to diffusion waves. A diffusion wave would carry finite mass and make a contribution to the conservation law (1.7), in addition to that of the translation of the detonation [2]. A weak detonation, on the other hand, is undercompressive and a perturbation would give rise to a diffusion wave. As a consequence, the time-asymptotic wave location cannot be determined a priori. Different analysis is needed for its stability. We leave this to the future.

2. Preliminaries. Travelling waves (1.1) and (1.4) satisfy

$$s' = \beta^{-1}(-\sigma(s + y) + f(s) + c) \equiv g(s, y), \tag{2.1}_1$$

$$y' = -\frac{K}{\sigma}\phi(s)y, \tag{2.1}_2$$

$$(s, y)(-\infty) = (u_l, 0), \qquad (s, y)(\infty) = (u_r, q).$$

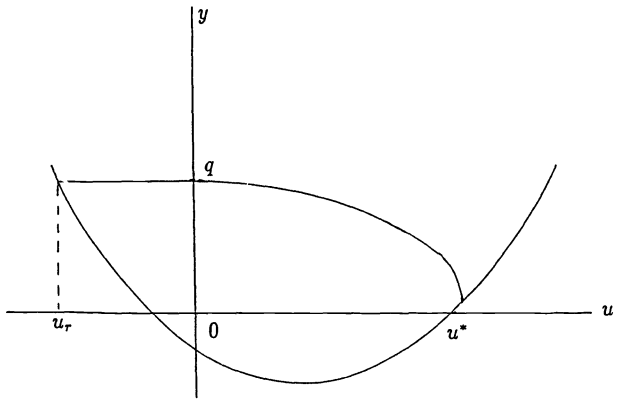


FIG. 2. *Phase diagram for strong detonations.*

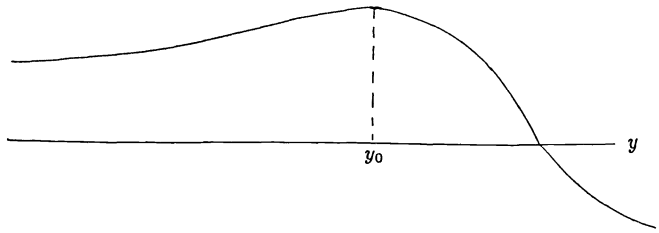


FIG. 3. *Fluid variable profile for strong detonation.*

We are interested in strong detonations, such as $u_l = u^*$. From the compressibility property (1.6), $(u^*, 0)$ is an unstable node for (2.1) and there is a unique connecting orbit provided that σ is not small [3], [4]. The trajectory is not monotone around $(u^*, 0)$ (see Fig. 2):

$$\begin{aligned} &f'(s)_x > 0 \quad \text{for } \eta < \eta_0, \\ &f'(s)_x < 0 \quad \text{for } \eta > \eta_0 \end{aligned} \tag{2.2}$$

for some constant η_0. For small heat release q, the effect of y in the $(2.1)_1$ is small and therefore s is almost monotone:

$$\begin{aligned} &0 < \int_{-\infty}^{\eta_0} f'(s)_\eta \, d\eta = \delta_1 << 1, \\ &f'(s, \eta) - \sigma > C \quad \text{for } -\infty < \eta < \eta_0 \end{aligned} \tag{2.3}$$

for some positive constant C; see Fig. 3.

3. Time-asymptotic analysis. Consider a perturbation of the strong detonation $(u^*, 0; u_r, 1)$:

$$(u, qz)(x, 0) = (s, y)(x) + (\bar{u}, \bar{y})(x, 0) \equiv (u_0, qz_0)(x). \tag{3.1}$$

From the compressibility property (1.6) there is no characteristic leaving the detonation and carrying diffusion waves. Thus the perturbation can cause, time-asymptotically, only the translation of the detonation. Suppose the amount of translation is x_0:

$$(u, qz)(x, t) \to (s, y)(x + x_0 - \sigma t) \quad \text{as } t \to \infty. \tag{3.2}$$

From (1.7), letting $t \to \infty$, we obtain

$$\int_{-\infty}^{\infty} \{(s+y)(x+x_0-\sigma t)-(s+y)(x-\sigma t)\}\,dx = \int_{-\infty}^{\infty} (\bar{u}+\bar{y})(x,0)\,dx.$$

Denote by $\psi(x_0)$ the left-hand side of the above identity; we have $\psi(0)=0$ and

$$\psi'(x_0) = \int_{-\infty}^{\infty} (s+y)'(x+x_0-\sigma t)\,dx = (s+y)|_{-\infty}^{\infty} = u_r + q - u^*.$$

It follows that

$$\psi(x_0) = x_0(u_r+q-u^*),$$
$$x_0 = (u_r+q-u^*)^{-1}\int_{-\infty}^{\infty} (\bar{u}+\bar{y})(x,0)\,dx.$$

This determines the time-asymptotic location of the detonation. We may translate (s,y) so that we can assume, without loss of generality, $x_0=0$, or equivalently,

$$\int_{-\infty}^{\infty} (\bar{u}+\bar{y})(x,t)\,dx = \int_{-\infty}^{\infty} (\bar{u}+\bar{y})(x,0) = 0, \qquad t \geq 0. \tag{3.3}$$

This is used by considering the antiderivative $v(x,t)$:

$$(u+qz)(x,t) = (s+y)(x-\sigma t) + v_x(x,t),$$
$$v(x,t) \equiv \int_{-\infty}^{x} (\bar{u}+\bar{y})(\xi,t)\,d\xi, \qquad v(\pm\infty,t)=0. \tag{3.4}$$

By integrating the difference of (1.1) and (2.1), we have

$$v_t + f(s+v_x+q(\zeta-z)) - f(s) = v_{xx} + q(\zeta-z)_x,$$
$$v(x,0) = \int_{-\infty}^{x} (\bar{u}+\bar{y})(\eta,0)\,d\eta, \tag{3.5}$$

(3.6)

$$(z-\zeta)(x,t) = z_0(x)\exp\left(-K\int_0^t \phi(u(x,\tau))\,d\tau\right) - \zeta(x)\exp\left(-K\int_0^t \phi(s(x,\tau))\,d\tau\right).$$

(3.5) can also be written as

$$v_t + f'(s)v_x = v_{xx} + q(\zeta-z)_x - f'(s)q(\zeta-z) + O(1)(v_x^2 + q^2(\zeta-z)^2). \tag{3.7}$$

4. Stability analysis. We start with the standard energy estimate by integrating (3.7) times v :

$$\frac{1}{2}\int_{-\infty}^{\infty} v^2(x,T)\,dx + \int_0^T\int_{-\infty}^{\infty}\left(v_x^2 - \frac{1}{2}f'(s)_x v^2\right)dx\,dt$$
$$= \frac{1}{2}\int_{-\infty}^{\infty} v^2(x,0)\,dx + \int_0^T\int_{-\infty}^{\infty}[-v_x q(\zeta-z) - vf'(s)q(\zeta-z)$$
$$+ O(1)v(v_x^2 + q^2(\zeta-z)^2)]\,dx\,dt.$$

With this and (2.2), we have from Cauchy–Schwarz that

$$\begin{aligned}\int_{-\infty}^{\infty} v^2(x,T)\,dx + \int_0^T\int_{-\infty}^{\infty} v_x^2\,dx\,dt + \int_0^T\int_{\sigma t+\eta_0}^{\infty} |f'(s)_x|v^2\,dx\,dt \\ \le \int_{-\infty}^{\infty} v^2(x,0)\,dx + \int_{-\infty}^{\infty} O(1)[q^2(\zeta-z)^2 + vq(\zeta-z)]\,dx\,dt \\ + \int_0^T\int_{-\infty}^{\sigma t+\eta_0} |f'(s)_x|v^2\,dx\,dt.\end{aligned} \tag{4.1}$$

Here we have made the following a priori assumption:

$$\sup_{-\infty<x<\infty,0\le t\le T} |v(x,t)| = \delta_2 << 1. \tag{4.2}$$

The second and third terms on the right-hand side of (4.1) are estimated, respectively, by using (3.6), (1.3) and (3.5), (1.6). We start with the latter by first rewriting (3.5) as

$$v_t + (f'(s)-\sigma)v_\eta = v_{\eta\eta} + q(\zeta-z)_\eta - f'(s)q(\zeta-z) + O(1)(v_\eta^2 + q^2(\zeta-z)^2),$$
$$\eta \equiv x - \sigma t,$$

and so

$$v_\eta = (f'(s)-\sigma)^{-1}[-v_t + v_{\eta\eta} + q(\zeta-z)_\eta - f'(s)\dot{q}(\zeta-z) + O(1)(v_\eta^2 + q^2(\zeta-z)^2)].$$

Integrating this times v over $(-\infty,\eta), \eta < \eta_0$, we have from (2.3) that

$$\begin{aligned}\frac{1}{2}v^2(\eta,t) &= \int_{-\infty}^{\eta} (f'(s)-\sigma)^{-1}[-vv_t + vv_{\eta\eta} + qv(\zeta-z)_\eta \\ &\qquad - vf'(s)q(\zeta-z) + O(1)v(v_\eta^2 + q^2(\zeta-z)^2)]d\eta \\ &= (f'(s)-\sigma)^{-1}(vv_\eta + qv(\zeta-z))(\eta,t) \\ &\quad + \int_{-\infty}^{\eta} (f'(s)-\sigma)^{-1}[-vv_t - v_\eta^2 - qv_\eta(\zeta-z) - vf'(s)q(\zeta-z) \\ &\quad + O(1)v(v_\eta^2 + q^2(\zeta-z)^2)]\,d\eta \\ &\quad + \int_{-\infty}^{\eta} (f'(s)-\sigma)^{-2}f'(s)_\eta[-vv_\eta - qv(\zeta-z)]\,d\eta,\end{aligned} \tag{4.3}$$

$$\begin{aligned}\int_0^T\int_{-\infty}^{\eta_0} f'(s)_\eta \frac{v^2}{2}\,d\eta\,dt \le \int_0^T\int_{-\infty}^{\eta_0} O(1)f'(s)_\eta(vv_\eta + qv(\zeta-z))\,d\eta\,dt \\ + O(1)\delta_1\int_0^T\int_{-\infty}^{\eta_0}[q^2(\zeta-z)^2 + O(1)vq(\zeta-z) \\ + (f'(s)_\eta v)^2]\,d\eta\,dt.\end{aligned}$$

This and (4.1) yield, for $\delta_1 + \delta_2$ small,

$$\begin{aligned}\int_{-\infty}^{\infty} v^2(x,T) + \int_0^T\int_{-\infty}^{\infty} v_x^2\,dx\,dt + \int_0^T\int_{-\infty}^{\infty} |f'(s)_x|v^2\,dx\,dt \\ \le \int_{-\infty}^{\infty} v^2(x,0)\,dx + \int_0^T\int_{-\infty}^{\infty}[O(1)q^2(\zeta-z)^2 + O(1)vq(\zeta-z)]\,dx\,dt.\end{aligned} \tag{4.4}$$

We now turn to the estimates for $\zeta - z$ using (3.6). From the hypotheses (4.2) and (1.3), there exists a positive constant D and $\eta_1, \eta_2, \eta_2 > \eta_1 > \eta_0$ such that

$$\phi(u) = 0 \quad \text{for } x - \sigma t > \eta_2, \tag{4.5}$$

$$\phi(u) > D \quad \text{for } x - \sigma t < \eta_1, \tag{4.6}$$

$$f'(s)_x < -D \quad \text{for } \eta_2 > x - \sigma t > \eta_1, \tag{4.7}$$

$$f'(s) - \sigma > D \quad \text{for } x - \sigma t < \eta_1; \tag{4.8}$$

cf. Fig. 3. Thus we have from (3.6) and (4.5) that

$$\begin{aligned} &(z-\zeta)(x,t) = z_0(x) - \zeta(x) \quad \text{for } x - \sigma t > \eta_2, \\ &\int_0^T \int_{\sigma t + \eta_2}^{\infty} [O(1)q^2(\zeta - z)^2 + O(1)vq(\zeta - z)]\, dx\, dt \\ &\qquad = O(1)q(q + \delta_2) \int_{\eta_2}^{\infty} |z_0(x) - \zeta(x)||x|\, dx. \end{aligned} \tag{4.9}$$

The region $\eta_1 < x - \sigma t < \eta_2$ has a finite width in both x and t directions and therefore we have from (3.6) and (3.4) that

$$\begin{aligned} (z - \zeta(x)) &= (z_0(x) - \zeta(x)) \exp\left(-K \int_{(x-\eta_2)/\sigma}^{t} \phi(u(x,\tau))\, d\tau\right) \\ &\quad + \zeta(x) \left[\exp\left(-K \int_{(x-\eta_2)/\sigma}^{t} \phi(s(x,\tau))\, d\tau\right) - \exp\left(-K \int_{(x-\eta_2)/\sigma}^{t} \phi(u(x,\tau))\, d\tau\right)\right] \\ &= O(1)(z_0(x) - \zeta(x)) + O(1) \int_{(x-\eta_2)/\sigma}^{t} (\phi(s(x,\tau)) - \phi(u(x,\tau)))\, d\tau \\ &= O(1)(z_0(x) - \zeta(x)) + \int_{(x-\eta_2)/\sigma}^{t} O(1)(q(z-\zeta) + v_x)(x,\tau)\, d\tau \\ &= O(1)(z_0(x) - \zeta(x)) + \left[O(1) \int_{(x-\eta_2)/\sigma}^{t} (q(z-\zeta) + v_x)^2(x,\tau)\, d\tau\right]^{1/2}, \end{aligned}$$

where we have used the Cauchy–Schwarz inequality. This and (4.7) yield

$$\begin{aligned} &\int_0^T \int_{\eta_1 + \sigma t}^{\eta_2 + \sigma t} [q^2(\zeta - z)^2 + |vq(\zeta - z)|]\, dx\, dt \\ &\quad = \int_0^T \int_{\eta_1 + \sigma t}^{\eta_2 + \sigma t} O(1)[q(\zeta - z)^2 + qv^2]\, dx\, dt \\ &\quad = \int_0^T \int_{\eta_1 + \sigma t}^{\eta^2 + \sigma t} O(1)q[(z_0(x) - \zeta(x))^2 + \int_{(x-\eta_2)/\sigma}^{t} (q(z-\zeta) + v_x)^2(x,\tau)\, d\tau]\, dx\, dt \\ &\qquad + \int_0^T \int_{\eta_1 + \sigma t}^{\eta_2 + \sigma t} O(1)q|f'(s)_x|v^2\, dx\, dt \\ &\quad = \int_{\eta_1}^{\infty} O(1)q|z_0(x) - \zeta(x)|^2\, dx \\ &\qquad + \int_0^T \int_{\eta_1 + \sigma t}^{\eta_2 + \sigma t} O(1)[q^2(z-\zeta)^2 + qv_x^2 + q|f'(s)_x|v^2]\, dx\, dt. \end{aligned}$$

Hence we have, for q small,

$$\begin{aligned}
&\int_0^T \int_{\eta_1+\sigma t}^{\eta_2+\sigma t} [O(1)q^2(\zeta - z)^2 + |vq(\zeta - z)|]\, dx\, dt \\
&= \int_{\eta_1}^{\infty} O(1)q|z_0(x) - \zeta(x)|^2\, dx \\
&\quad + \int_0^T \int_{\eta_1+\sigma t}^{\eta_2+\sigma t} O(1)q(v_x^2 + |f'(s)_x|v^2)\, dx\, dt.
\end{aligned} \tag{4.10}$$

We finally turn to the same integral as (4.9) and (4.10) but for the region $x - \sigma t < \eta_1$ by using (4.6) and (3.6)

$$\begin{aligned}
(z-\zeta)(x,t) &= O(1)\left[|z_0(x) - \zeta(x)| + \int_0^t |\phi(s(x,\tau)) - \phi(u(x,\tau))|\, d\tau\right] \\
&\quad \times \exp\left(-K\int_0^t \phi(s(x,\tau))\, d\tau\right) \\
&= O(1)[|z_0(x) - \zeta(x)| \\
&\qquad + \int_{t_1(x)}^t |\phi(s(x,\tau)) - \phi(u(x,\tau))| d\tau] \exp(-KDt_2(x)), \\
t_1(x) &\equiv \max((x-\eta_2)/\sigma, 0), t_2(x) \equiv \max((x-\eta_1)/\sigma, 0).
\end{aligned} \tag{4.11}$$

In particular, we have for some $C > 0$,

$$(z-\zeta)(x,t) = O(1)(z_0(x) + \zeta(x))\exp(-Ct) \quad \text{for } x < 0.$$

Thus

$$\int_0^T \int_{-\infty}^0 [q^2(z-\zeta)^2 + |vq(\zeta - z)|]\, dx\, dt = O(1)q(q+\delta_2)\int_{-\infty}^0 (z_0(x) + \zeta(x))\, dx. \tag{4.12}$$

Again from (4.11), for $x > 0$ and $\eta = x - \sigma t < \eta_1$,

$$(z-\zeta)(x,t) = O(1)\left[|z_0(x) - \zeta(x)| + \int_{t_1(x)}^t |\phi(s(x,\tau)) - \phi(u(x,\tau))| d\tau\right] \exp(-C|\eta - \eta_1|),$$

whence

$$\begin{aligned}
&\int_{|\eta_1|/\sigma}^T \int_0^{\eta_1+\sigma t} [q^2(z-\zeta)^2 + |vq(\zeta - z)|]\, dx\, dt \\
&\le q\int_0^T \int_0^{\sigma t} O(1)|z_0(x) - \zeta(x)|^2 e^{-Ct}\, dx\, dt \\
&\quad + q\int_{|\eta_1|/\sigma}^T \int_0^{\eta_1+\sigma t} O(1)\left[\int_{t_1(x)}^t |s(x,\tau) - u(x,\tau)|\, d\tau\right]^2 \\
&\quad \times e^{-C|x-\sigma t|}\, dx\, dt + q\int_{|\eta_1|/\sigma}^T \int_0^{\eta_1+\sigma t} O(1)v^2 \times e^{-C|x-\sigma t|}\, dx\, dt \equiv \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

Clearly,

$$\mathrm{I} = O(1)q\int_0^\infty |z_0(x)-\zeta(x)|^2\,dx$$

and, by a change of order of integrations,

$$\mathrm{III} \le O(1)q\int_0^T\int_{-\infty}^{\eta_1} O(1)v^2(\eta,t)e^{-C|\eta|}\,d\eta\,dt, \qquad \eta \equiv x-\sigma t.$$

Thus we may apply the same procedure in deriving (4.4) from (4.3) to obtain

$$\begin{aligned}\mathrm{III} &= O(1)q\int_0^T\int_{-\infty}^{\eta_1}[q^2(\zeta-z)^2+O(1)vq(\zeta-z)+f'(s)_\eta v^2]\,dx\,dt\\ &= O(1)q\int_0^T\int_{-\infty}^{\eta_1+\sigma t}[q^2(\zeta-z)^2+O(1)|vq(\zeta-z)|+|f'(s)_\eta v^2|]\,dx\,dt.\end{aligned}$$

From (3.4) and Cauchy–Schwarz we see that

$$\begin{aligned}\mathrm{II} &\le q\int_0^T\int_0^{\sigma t} O(1)\left[\int_{(x-\eta_2)/\sigma}^t (|q(z-\zeta)|+|v_x|)(x,\tau)\,d\tau\right]^2 e^{-C|x-\sigma t|}\,dx\,dt\\ &= q\int_0^T\int_0^{\sigma t} O(1)\left[\int_{(x-\eta_2)/\sigma}^t (q^2(z-\zeta)^2+v_x^2)(x,\tau)\,d\tau\right]|x-\sigma t|e^{-C|x-\sigma t|}\,dx\,dt\\ &= O(1)q\int_0^{\sigma t}\int_{x/\sigma}^T\left[\int_{(x-\eta_2)/\sigma}^t (q^2(z-\zeta)^2+v_x^2)(x,\tau)\,d\tau\right]e^{-C|x-\sigma t|/2}\,dx\,dt\\ &= O(1)q\int_0^{\sigma t}\int_{(x-\eta_2)/\sigma}^T (q^2(z-\zeta)^2+v_x^2)\,dx\,dt.\end{aligned}$$

We conclude from these estimates for I–III and (4.9)–(4.12) that, for q small,

$$\begin{aligned}&\int_0^T\int_{-\infty}^\infty [O(1)q^2(\zeta-z)^2+O(1)qv(\zeta-z)]\,dx\,dt\\ &\quad= O(1)q\int_0^\infty x|z_0(x)-\zeta(x)|\,dx+O(1)q\int_{-\infty}^0 (z_0(x)+\zeta(x))\,dx\\ &\qquad+O(1)q\int_0^T\int_{-\infty}^\infty (|f'(s)_x|v^2+v_x^2)\,dx\,dt.\end{aligned}\tag{4.13}$$

Finally, we have from (4.4) and (4.13), for $q+\delta_1+\delta_2$ sufficiently small, the following main energy estimate:

$$\begin{aligned}&\int_{-\infty}^\infty v^2(x,T)+\int_0^T\int_{-\infty}^\infty (v_x^2+|f'(s)_x|v^2)\,dx\,dt\\ &\quad\le \int_{-\infty}^\infty v^2(x,0)\,dx+O(1)q\int_0^\infty x|z_0(x)-\zeta(x)|\,dx\\ &\qquad+O(1)q\int_{-\infty}^0 (z_0(x)+\zeta(x))\,dx.\end{aligned}\tag{4.14}$$

We now state our stability result.

THEOREM. *Suppose that q is small,*

$$\int_{-\infty}^{x} (u_0(x) - s(x))\,dx \quad \text{and} \quad \int_{-\infty}^{x} (z_0(x) - \zeta(x))\,dx$$

are sufficiently small in the space $H^3(-\infty, +\infty)$, and that as $x \to \pm\infty$,

$$u_0(x) - s(x) = O(|x|^{-3/2-\epsilon}), \qquad u_0'(x) - s'(x) = O(|x|^{-1/2-\epsilon}),$$
$$u_0''(x) - s''(x) = O(|x|^{-1/2-\epsilon}) \qquad \frac{d^k}{dx^k}(z_0(x) - \zeta(x)) = O(|x|^{-2-\epsilon}),$$

$k = 0, \ldots, 3$, for some $\epsilon > 0$. Then $u(\cdot, t) - s(\cdot - \sigma t)$ and $z(\cdot, t) - \zeta(\cdot - \sigma t)$ decay to zero in the $L^\infty(x)$ as $t \to +\infty$.

Proof. From our hypotheses, the energy estimate (4.14) becomes

$$\int_{-\infty}^{\infty} v^2(x, T) + \int_0^T \int_{-\infty}^{\infty} (v_x^2 + |f'(s)_x| v^2)\,dx\,dt = O(1)\delta,$$

for a small constant δ. By similar arguments we may use the differentiations of (3.4) and (3.5) to obtain energy estimates for higher derivatives of v :

$$\int_{-\infty}^{\infty} w^2(x, T)\,dx + \int_0^T \int_{-\infty}^{\infty} \left(\frac{\partial w}{\partial x}\right)^2 dx\,dt = O(1)\delta,$$
$$\int_{-\infty}^{\infty} w_1^2(x, 0)\,dx + \int_0^T \int_{-\infty}^{\infty} \left(\frac{\partial w_1}{\partial x}\right)^2 dx\,dt = O(1)\delta,$$
$$w \equiv v_x = u + qz - (s + q\zeta), \qquad w_1 \equiv w_x.$$

In turn these estimates imply that the assumption (3.5) holds by the Sobolev inequality. In fact, the same arguments also yield the equicontinuity of the L^2 norms of w and w_1. Since their double integrals are bounded, it follows that

$$\int_{-\infty}^{\infty} w^2(x, T)\,dx \to 0 \quad \text{as } T \to \infty.$$

To finish the proof, it remains to show that one of the quantities, $(u - s)(x, t)$ and $(z - \zeta)(x, t)$, goes to zero as $t \to \infty$.

From (3.6)

$$\begin{aligned}(u - s)(x, t) &= w(x, t) - q(z - \zeta)(x, t)\\ &= w(x, t) - q(z_0(x) - \zeta(x)) \exp\left(-K \int_0^t \phi(u(x, \tau))\,d\tau\right)\\ &\quad - q\zeta(x)\left[\exp\left(-K \int_0^t \phi(u(x, \tau))\,d\tau\right) - \exp\left(-K \int_0^t \phi(s(x, \tau))\,d\tau\right)\right]\\ &= w(x, t) + O(1)q(z_0(x) - \zeta(x)) + Kq\zeta(x) \int_0^t O(1)(u - s)(x, \tau)\,d\tau.\end{aligned}$$

Take a constant $E, \sigma > E > 0$. We have for $x > Et$,

$$O(1)q(z_0(x) - \zeta(x)) = O(t^{-2-\epsilon}) \to 0.$$

Thus it follows from Gronwall's inequality that $(u-s)(x,t) \to 0$ as $t \to 0$. For $x < Et$, we have from (3.6) and (4.6) that

$$\begin{aligned}(z-\zeta)(x,t) &= O(1)(z_0+\zeta)(x)\exp\left(-KD\left(t-\frac{x-\eta_1}{\sigma}\right)\right)\\ &= O(1)\exp(-t[1-D/\sigma]+\eta_1/\sigma) \to 0 \quad \text{as } t \to 0.\end{aligned}$$

This completes the proof of the theorem.

REFERENCES

[1] W. FRICKET AND W. C. DAVIS, *Detonation*, University of California Press, Berkeley, CA, 1979.

[2] T.-P. LIU, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc., 328, American Mathematical Society, Providence, RI, 1985.

[3] A. MAJDA, *A qualitative model for dynamic combustion*, SIAM J. Appl. Math., 41 (1981), pp. 70–91.

[4] Z. H. TENG AND L. A. YING, *Existence, uniqueness and convergence as vanishing viscosity for a reaction-diffusion-convection system*, Acta Math. Sinica (N.S.), 2 (1989), pp. 114–135.

[5] L. A. YING AND Z. H. TENG, *Riemann problem for a reaction and convection hyperbolic system*, Approx. Theory Appl., 1 (1984), pp. 95–122.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 529–565, May 1995

002

APPROXIMATION PRÈS DU TEMPS D'EXPLOSION DES SOLUTIONS D'ÉQUATIONS D'ONDE QUASI-LINÉAIRES EN DIMENSION DEUX*

SERGE ALINHAC†

Abstract. For a general quasi-linear wave equation in two space dimensions and Cauchy data of size ϵ, we construct an approximate solution using the method of nonlinear geometric optics.

Away from the blow up time, we obtain arbitrary accuracy near the light cone.

Near the blow up time, we make explicit the behavior of the solution and of the error terms.

Key words. équation d'onde quasi-linéaire, solution approchée, optique géométrique non linéaire

AMS subject classification. 35L40

Introduction. Dans ce travail, nous étudions les solutions d'équations d'onde quasi-linéaires en dimension deux d'espace, pour des données de Cauchy C_0^∞ de taille ε.

De nombreux mathématiciens ont contribué à cette étude, parmi lesquels on peut citer Klainerman, John, et Hörmander (on trouvera dans [6] une bibliographie détaillée).

Il a été notamment établi que le temps de vie T_ε de la solution est de l'ordre de A_0^2/ε^2, et que la solution u est approximée, pour tout $A < A_0$ et $t \leq A^2/\varepsilon^2$, par une fonction u_a de la forme

$$u_a = \frac{\varepsilon}{r^{1/2}} R(r-t, \omega, \tau), \quad \text{où } x = r(\cos\omega, \sin\omega), \qquad \tau = \varepsilon\sqrt{t},$$

R étant solution d'une équation de type Burger.

Ces procédés d'approximations, qualifiés "d'optique géométrique non linéaire" sont classiques (voir, par exemple, Di Perna et Majda [3], Majda [8], [9], Hörmander [5], ou encore [1]).

Le présent article a pour but de décrire une construction beaucoup plus précise, en insistant sur le comportement de la solution u_a et de l'erreur près du bord du cône de lumière, c'est-à-dire dans une zone de la forme $-\text{cte} \leq r - t \leq M$ (les données de Cauchy étant supportées dans $|x| \leq M$).

De plus, on ne se limitera pas à $t \leq A^2/\varepsilon^2$ $(A < A_0)$: on établira les comportements de la solution et de l'erreur pour

$$t \leq \frac{A_0^2}{\varepsilon^2} + 2\frac{A_0 A_1}{\varepsilon} - o\Big(\frac{1}{\varepsilon}\Big),$$

c'est-à-dire au voisinage du temps d'explosion de $\nabla^2 u_a$.

La motivation de cette étude est le désir de préciser les estimations connues de T_ε et de comprendre la nature du comportement explosif de u lorsque $t \underset{<}{\to} T_\varepsilon$. Un premier pas dans cette direction a été fait dans [2].

---

*Received by the editors February 16, 1993; accepted for publication (in revised form) September 23, 1993.

†Université de Paris-Sud, Département de Mathématiques, 91405 Orsay Cedex, France.

Cet article étant par nécessité assez technique, nous avons adopté le plan suivant.

• Au §1 sont introduites les notations et le problème précis.

• Au §2, nous résumons la construction et les estimations d'erreurs zone par zone, après quelques commentaires destinés à aider le lecteur à saisir les enjeux.

• Le §3 est consacré à une première période de temps où les interactions non linéaires sont faibles.

• Au §4, on explicite la transition assez délicate de cette première période au temps $t \leq A^2/\varepsilon^2$ $(A < A_0)$, en résumant les résultats obtenus au §5.

• Enfin, le §6 est consacré à l'étude pour $t \geq A^2/\varepsilon^2$ et jusqu'à une distance $o\left(\frac{1}{\varepsilon}\right)$ du temps d'explosion.

1. Notations et position du problème.

• Dans R^3, on note (x_0, x_1, x_2) les variables, en utilisant souvent la notation commode $x_0 = t$, $x = (x_1, x_2)$.

Les coordonnées polaires en (x_1, x_2) seront alors notées (r, ω), avec $r = (x_1^2 + x_2^2)^{1/2}$, $x_1 = r\cos\omega$, $x_2 = r\sin\omega$, $\omega_1 = \cos\omega$, $\omega_2 = \sin\omega$, $\partial_\omega = x_1\partial_2 - x_2\partial_1$.

Pour une fonction $f(x,t)$, on notera par abus

$$|f|_0 = \left(\int |f(x,t)|^2\, dx\right)^{1/2} \quad \text{et} \quad \|f\|_0 = \sup_x |f(x,t)|\,,$$

les normes L^2 et L^∞ de f à t fixé.

• On considère l'équation des ondes quasi-linéaire à coefficients réels

$$\partial_t^2 u - \Delta_x u + g_{ij}^k\, \partial_k u\, \partial_{ij}^2 u = 0, \tag{1.1}$$

où la sommation est étendue aux indices $0 \leq i, j, k \leq 2$, $\partial_k = \partial/\partial x_k$, $g_{ij}^k = g_{ji}^k$, $g_{00}^k = 0$, $\Delta_x = \partial_1^2 + \partial_2^2$ et $\Box = \partial_t^2 - \Delta_x$.

Avec $\omega_0 = -1$, on définit comme dans [5]

$$g(\omega) = g_{ij}^k\, \omega_i \omega_j \omega_k. \tag{1.2}$$

On suppose données des fonctions $u^0(x,\varepsilon)$, $u^1(x,\varepsilon)$, réelles de classe C^∞ dans $\mathrm{R}^n \times [0, \varepsilon_0[$, supportées dans $|x| \leq M$, pour lesquelles on a

$$u^0(x,\varepsilon) = \varepsilon u_1^0(x) + \varepsilon^2 u_2^0(x) + \cdots,\ u^1(x,\varepsilon) = \varepsilon u_1^1(x) + \varepsilon^2 u_2^1(x) + \cdots.$$

On considère, pour $\varepsilon > 0$, le problème de Cauchy

$$\Box u + g_{ij}^k\, \partial_k u\, \partial_{ij}^2 u = 0, \quad u(x,0) = u^0(x,\varepsilon), \quad \partial_t u(x,0) = u^1(x,\varepsilon). \tag{1.3}$$

On se propose de construire une solution approchée u_a du problème (1.3), en portant une attention particulière aux comportements de la solution u_a et de l'erreur $J_a = \Box u_a + g_{ij}^k\, \partial_k u_a\, \partial_{ij}^2 u_a$ près du temps d'explosion.

2. Résultats obtenus: Description de la solution approchée et estimations des erreurs.

2.1. Quelques généralités.

2.1.1. Dans la description de la solution approchée u_a de (1.3) qui sera construite en détail aux §§3–6, nous distinguons systématiquement deux zones d'espace temps et trois périodes de temps.

• La zone dite "extérieure" est située au bord du cône de lumière, et est définie par $-C_0 \leq r - t \leq M$, pour une constante C_0 assez grande (en fait, $C_0 \gg |\sigma_0|$, voir plus bas).

• La zone dite "intérieure," définie par $r - t \leq -C_0 + 3$. Les trois périodes de temps sont définies comme suit.

• La "période I" est $0 \leq t \leq 2\varepsilon^{-\lambda}$, où $\lambda < 2$ est à choisir (on prendra en fait $\lambda = \frac{14}{9}$).

• La "période II" est $\varepsilon^{-\lambda} \leq t \leq A^2/\varepsilon^2$, où A est un nombre positif fixé arbitrairement, avec toutefois $A < A_0$ (voir plus bas).

• La "période III" commence à $t = A^2/\varepsilon^2$, et s'achève avec l'explosion de la solution.

La solution approchée u_a sera en fait construite séparément dans chaque domaine, les fonctions correspondantes étant notées $u_a^{I,i}$ (i pour "intérieur"), $u_a^{I,e}$ (e pour "extérieur"), $u_a^{II,i}$, etc.

La solution u_a sera obtenue enfin par recollement des morceaux à l'aide de troncatures.

• Pour $\theta \in C^\infty(\mathrm{R})$ vérifiant $\theta(s) = 1$ pour $s \leq 1$ et $\theta(s) = 0$ pour $s \geq 2$, on pose

$$u_a^I = (1-\theta)(r-t+C_0-1)u_a^{I,e} + \theta(r-t+C_0-1)u_a^{I,i} \ ,$$

et de même pour les périodes II et III.

• Pour les périodes I et II, on pose

$$u_a = \theta(te^\lambda)u_a^I + (1-\theta(te^\lambda))u_a^{II} \ .$$

• Dans un souci de simplification, le passage de la période II à la période III se fait par simple recollement de u_a^{II} et u_a^{III} sur $t = A^2/\varepsilon^2$.

2.1.2. La constante C_0 choisie en §2.1.1 est assez grande pour assurer le fait suivant : la solution approchée u_a construite n'explose pas en zone intérieure, même en période III.

C'est en zone extérieure que se produit (en période III) l'explosion de la solution approchée u_a.

Du fait qu'on espère voir la vraie solution u se comporter comme u_a, on s'intéresse donc tout spécialement à la zone extérieure.

La preuve que u manifeste effectivement un "comportement explosif" près des points où u_a explose est délicate, et nécessite que u_a soit une très bonne approximation de u (cette preuve est fournie en [2]). C'est ce point qui explique la différence de traitement du présent travail entre les deux zones.

La transition "vers l'intérieur" s'opérera donc essentiellement en négligeant dans u_a^e des termes devenus inutiles pour u_a^i.

En revanche, l'absence de lacune pour les solutions libres de l'équation des ondes en dimension deux, et les interactions non linéaires, obligent à considérer le comportement de u_a^i pour de grandes valeurs de $t - r$. Or l'on sait [5] qu'une solution u à données C_0^∞ de $\Box u = 0$ s'écrit

$$u = \frac{1}{r^{1/2}} F\Big(r-t, \omega, \frac{1}{r}\Big),$$

où $F(\sigma,\omega,z)$ vérifie des inégalités "de type symbole"

$$\left|\partial_\omega^\alpha \partial_z^\beta \partial_\sigma^\gamma F(\sigma,\omega,z)\right| \leq C_{\alpha\beta\gamma}(1+|\sigma|)^{-\frac{1}{2}+|\beta|-|\gamma|}.$$

Si l'on développe F par la formule de Taylor en $z=0$, on obtient

$$u \sim \frac{1}{r^{1/2}}\Big\{F_0(r-t,\omega)+\frac{1}{r}F_1(r-t,\omega)+\cdots+\frac{1}{r^k}F_k(r-t,\omega)+\cdots\Big\},$$

où $F_k(\sigma,\omega)$ est un symbole en σ d'ordre $-\frac{1}{2}+k$ (et de type $(1,0)$).

Ce développement ne nous apprend donc rien sur le comportement de u dans toute zone $\varepsilon \leq \dfrac{r}{t} \leq 1-\varepsilon$ $(\varepsilon>0)$, tandis qu'il est arbitrairement précis en zone extérieure.

On voit ainsi que les termes négligés de u_a^e seraient devenus dans u_a^i, de toutes façons, non significatifs.

2.1.3. Les formes successives de u_a à travers les trois périodes se laissent comprendre de la façon suivante.

En période I, on calcule tout simplement u_a par un développement en série de ε

$$u_a = \varepsilon u_1 + \varepsilon^2 u_2 + \cdots$$

• Les interactions quadratiques "de première génération" (c'est-à-dire celles de

$$u_1 \sim \frac{S_0(r-t,\omega,\frac{1}{r})}{r^{1/2}}$$

avec lui-même) font apparaître des termes en

$$\varepsilon^2\sqrt{t}\,\frac{1}{r^{1/2}}S_1\Big(r-t,\omega,\frac{1}{r}\Big), \quad \text{puis } \varepsilon^3 t\,\frac{1}{r^{1/2}}S_2\Big(r-t,\omega,\frac{1}{r}\Big), \text{etc.}$$

les S_i étant certains symboles en σ. Cela conduit à introduire le temps lent $\tau=\varepsilon\sqrt{t}$, de façon à écrire la somme

$$\frac{\varepsilon}{r^{1/2}}(S_0+\tau S_1+\tau^2 S_2+\ldots)$$

sous la forme

$$\frac{\varepsilon}{r^{1/2}}S\Big(r-t,\omega,\frac{1}{r},\tau\Big).$$

• Les interactions de "deuxième génération" (c'est-à-dire celles de u_1 avec u_2) font apparaître, outre des termes comme ci-dessus, des termes de la forme

$$\frac{\varepsilon^3\log t}{r^{1/2}}S_3\Big(r-t,\omega,\frac{1}{r}\Big), \quad \text{puis } \frac{\varepsilon^5(\log t)^2}{r^{1/2}}S_3\Big(r-t,\omega,\frac{1}{r}\Big), \textit{etc.}$$

Cela conduit à introduire un deuxième temps lent $\zeta=\varepsilon^2\log t$, tous les termes se sommant en

$$\frac{\varepsilon}{r^{1/2}}F\Big(r-t,\omega,\frac{1}{r},\tau,\zeta\Big).$$

En période II, on cherchera donc la solution sous la forme modulée

$$u_a \sim \frac{\varepsilon}{r^{1/2}} F\Big(r-t, \omega, \frac{1}{r}, \tau, \zeta\Big),$$

l'équation aux dérivées partielles satisfaite par F traduisant les effets des interactions quadratiques.

En première approximation, (et notamment, en zone intérieure), cette équation n'est autre que l'équation de Burger en les variables $\sigma = r - t$ (espace) et τ (temps). C'est un fait classique (cf., par exemple, [5]).

Si l'on recherche une précision supérieure, on doit faire appel en deuxième période à un modèle plus complexe, l'apparition des termes en $(\log t)^k$ décrits plus haut correspondant à des valeurs entières d'indices pour certaines équations à points singuliers-réguliers (cf. [10]).

La transition entre les périodes I et II consistera à imposer à F comme conditions initiales sur $\tau = \zeta = 0$ les valeurs obtenues en période I.

En période III, le temps lent ζ et la variable $z = \dfrac{1}{r}$ deviennent redondants, car

$$z = \frac{\varepsilon^2}{\tau^2}\Big(1 + \frac{\sigma\varepsilon^2}{\tau^2}\Big)^{-1}, \qquad \zeta = 2\varepsilon^2 \log \frac{\tau}{\varepsilon}.$$

On va donc simplifier, pour $\tau \geq A$, la forme de u_a en $u \sim \dfrac{\varepsilon}{r^{1/2}} F_\varepsilon(r-t, \omega, \tau)$ (cette nouvelle fonction F_ε dépendant alors de ε). Le raccord entre les zones II et III consistera à imposer

$$F_\varepsilon(\sigma, \omega, A) = F\Big(\sigma, \omega, \frac{\varepsilon^2}{A^2}\Big(1 + \frac{\sigma\varepsilon^2}{A^2}\Big)^{-1}, A, 2\varepsilon^2 \log \frac{A}{\varepsilon}\Big).$$

Bien entendu, dans l'étude de $u_a^{III,e}$, on porte une attention particulière à l'explosion des dérivées d'ordre supérieur ou égal à deux. Ce sont ces résultats qui servent de point de départ, dans [2], à l'étude du comportement explosif de la vraie solution u.

2.1.4. L'erreur liée à l'approximation de u par u_a sera évaluée en calculant le comportement de $J_a = \Box u_a + g_{ij}^k \partial_k u_a \partial_{ij}^2 u_a$, et en obtenant des estimations de $\dot{u} = u - u_a$ en périodes I et II.

2.2. Description en période I. On définit les fonctions $u_j(x,t)$ $(j \geq 1)$ par

$$\Box u_1 = 0, \quad u_1(x,0) = u_1^0(x), \quad \partial_t u_1(x,0) = u_1^1(x), \tag{2.2.1}$$

$$\Box u_j + Q_j = 0, \qquad u_j(x,0) = u_j^0(x), \partial_t u_j(x,0) = u_j^1(x) \quad \text{pour } j \geq 2, \tag{2.2.2}$$

où $Q_p = \displaystyle\sum_{\ell+\ell'=p} g_{ij}^k \partial_k u_\ell \partial_{ij}^2 u_{\ell'}$.

Il est bien connu (cf. [4], [5]) que

$$u_1 \sim \frac{R(r-t,\omega)}{r^{1/2}} \quad (r \to +\infty, r - t \geq -C), \tag{2.2.3}$$

pour

$$R(\sigma,\omega)=\frac{1}{2\sqrt{2\pi}}\int_{s\geq\sigma}\frac{1}{\sqrt{s-\sigma}}\Big\{\tilde{R}(s,\omega,u_1^1)-\partial_s\tilde{R}(s,\omega,u_1^0)\Big\}\,ds, \tag{2.2.4}$$

$\tilde{R}(s,\omega,v)$ désignant la transformée de Radon

$$\tilde{R}(s,\omega,v)=\int_{x\omega=s}v(x)\,dx \quad \text{de } v.$$

Par ailleurs, on établit au lemme 3.2.1 que

$$u_2-\frac{g(\omega)}{2}(\partial_\sigma R)^2\sim\frac{L(r-t,\omega)}{r^{1/2}}\quad (r\to+\infty,\ r-t\geq -C)\ , \tag{2.2.5}$$

pour une certaine fonction C^∞ $L(\sigma,\omega)$.

2.2.1. En zone extérieure, la proposition 3.1.1 fournit une approximation arbitraire de u_k : pour tous $N,N'\in\mathbb{N}$, il existe des fonctions C^∞ L_k^ℓ et $R_k^{\ell,\ell'}$ telles que

$$u_k=\frac{1}{r^{1/2}}\Bigg\{\sum_{0\leq 2\ell\leq k-1}(\log t)^\ell L_k^\ell\Big(r-t,\omega,\frac{1}{r}\Big)+\sum_{\substack{\ell\geq 1\\ \ell+2\ell'\leq k-1}}t^{\ell/2}(\log t)^{\ell'}R_k^{\ell,\ell'}\Big(r-t,\omega,\frac{1}{r}\Big)\Bigg\}+r_k,$$

avec $r_k=O(r^{-N})$ et $\Box L_k^\ell/r^{1/2}=O(r^{-N'})$.

On considérera dans la suite N et N' comme fixés, très grands, selon les besoins. Les dépendances de L_k^ℓ et $R_k^{\ell,\ell'}$ ne sont jamais explicitées.

On prendra, pour q grand, à choisir,

$$u_a^{I,e}=\sum_{\ell=1}^{q}\varepsilon^\ell u_\ell, \tag{2.2.6}$$

d'où il résulte par construction $J_a=O(\varepsilon^{q+1}t^{(q-3)/2})$.

2.2.2. En zone intérieure, nous n'utilisons les termes u_k que pour $k\leq 7$. Nous prenons d'abord $u_1^i=u_1$, $u_2^i=u_2$; pour u_3^i et u_4^i, nous ne gardons que les deux termes principaux de $u_{3,a}$ et $u_{4,a}$; enfin, pour u_5^i, u_6^i et u_7^i, nous ne gardons que les termes principaux de $u_{5,a}$, $u_{6,a}$ et $u_{7,a}$ (tous ces termes sont explicités en §3.2).

Nous posons donc

$$u_a^{I,i}=\sum_{\ell=1}^{7}\varepsilon^\ell u_\ell^i\ . \tag{2.2.7}$$

2.2.3. Après recollement, on obtient pour J_a les estimations

$$\big|\partial_{x,t,\omega}^\alpha J_a\big|_0\leq C_\alpha\Big\{\frac{\varepsilon^3}{1+t}+\varepsilon^5(1+t)^{1/2}+\varepsilon^8(1+t)^{5/2}\Big\}\,|\log t|^{q-1}\,, \tag{2.2.8}$$

avec de plus à l'extérieur

$$\|\partial^{\alpha}_{x,t,\omega} J_a\|_0 \leq C_\alpha \ \varepsilon^{q+1} \ t^{(q-3)/2}. \tag{2.2.9}$$

2.2.4. Les deux fonctions R et L jouent dans la suite un rôle essentiel. Introduisons dès maintenant quelques commentaires et hypothèses qui s'y rapportent.

• Nous faisons sur R et g l'hypothèse de non dégénérescence suivante :

(ND) Il existe un point (σ_0, ω_0) et un nombre $\kappa \geq 2$ tels que

(i) $-g(\omega_0)\,\partial^2_\sigma R(\sigma_0, \omega_0) < 0$,

(ii) $\forall A$, $\exists C > 0$ avec, pour $|\sigma - \sigma_0| + |\omega - \omega_0| \leq A$,

$$-g(\omega)\,\partial^2_\sigma R(\sigma, \omega) \geq -g(\omega_0)\,\partial^2_\sigma R(\sigma_0, \omega_0) + C(|\sigma - \sigma_0| + |\omega - \omega_0|)^\kappa.$$

Remarquons que cela implique que (σ_0, ω_0) est un point (unique) de minimum absolu de $-g(\omega)\,\partial^2_\sigma R(\sigma, \omega)$.

Dans le cas spécial où l'équation considérée serait de la forme $\partial_t^2 u - c^2(u_t)\,\Delta u = 0$ et les données initiales invariantes par rotation, cette hypothèse devrait être remplacée par

(ND)′ Il existe un point σ_0 et un nombre $\kappa \geq 2$ tels que

(i) $-g\,\partial^2_\sigma R(\sigma_0) < 0$,

(ii) $\forall A$, $\exists C > 0$ avec, pour $|\sigma - \sigma_0| \leq A$,

$$-g\,\partial^2_\sigma R(\sigma) \geq -g\,\partial^2_\sigma R(\sigma_0) + C\,|\sigma - \sigma_0|^\kappa \ .$$

• Définissons la fonction $S(\sigma, \omega, \tau)$ par

$$\partial_\tau S - \frac{g}{2}(\partial_\sigma S)^2 = 0, \quad S(\sigma, \omega, 0) = R(\sigma, \omega) + \varepsilon L(\sigma, \omega) \ . \tag{2.2.10}$$

La fonction $\varepsilon/r^{1/2}\, S(r - t, \omega, \tau)$ apparaîtra ultérieurement comme une assez bonne approximation de u.

Comme c'est la dérivée S' de S qui satisfait une équation de Burger, on retrouve bien ici le fait que ce sont les dérivées secondes $\nabla^2 u$ qui "doivent" exploser.

• Posons

$$A_0 = \frac{1}{g(\omega_0)\,\partial^2_\sigma R(\sigma_0, \omega_0)}, \quad A_1 = -A_0^2\, g(\omega_0)\,\partial^2_\sigma L(\sigma_0, \omega_0), \quad \text{et} \quad \tau_* = \tau_*(\varepsilon) = A_0 + \varepsilon A_1.$$

Compte tenu de l'hypothèse (ND) et des propriétés bien connues de l'équation de Burger, on note que A_0 est le temps de vie de S pour $\varepsilon = 0$, tandis que $\tau_*(\varepsilon)$ approxime à $0(\varepsilon^2)$ près le temps de vie de S dans le cas général.

2.3. Description en période *II*.

2.3.1. En zone extérieure, et compte tenu de la proposition 3.1.1, nous remarquons que, formellement,

$$u_a^{I,e} \sim \frac{\varepsilon}{r^{1/2}} F(r - t, \omega, \frac{1}{r}, \tau, \zeta).$$

La proposition 4.1.1 permet de construire, dans un intervalle $0 \leq \tau \leq \tau_1$ (où $\tau_1 = A_0 + o(1)$), une fonction $F(\sigma, \omega, z, \tau, \zeta)$ qui se raccorde bien à $u_a^{I,e}$. Nous posons donc

$$u_a^{II,e} = \frac{\varepsilon}{r^{1/2}} F\Big(r - t, \omega, \frac{1}{r}, \tau, \zeta\Big). \tag{2.3.1}$$

2.3.2. En zone intérieure, nous choisissons

$$u_a^{II,i} = \varepsilon u_1 + \varepsilon^2 \Big(u_2 - \frac{g}{2}\chi\Big(\frac{t}{r}\Big)^{1/2} (\partial_\sigma R)^2\Big) + \chi \frac{\varepsilon}{r^{1/2}} \Big[S(r-t, \omega, \tau) - S(r-t, \omega, 0)\Big] , \tag{2.3.2}$$

où u_1, u_2, R et S sont définis au §2.2, tandis que χ dénote (par abus) une troncature

$$\chi\Big(\frac{r}{1+t}\Big), \quad \text{avec } \chi \in C^\infty(R), \quad \chi(s) = 0 \quad \text{pour } s \leq \frac{1}{2}, \quad \chi(s) = 1 \quad \text{pour } s \geq \frac{2}{3}.$$

On choisit l'expression (2.3.2) pour des raisons techniques : le lemme 4.2.1 indique en fait l'équivalence

$$u_a^{II,i} \sim \chi \frac{\varepsilon}{r^{1/2}} S(r - t, \omega, \tau). \tag{2.3.3}$$

2.3.3. Après recollement spatial et temporel, on obtient pour J_a en période II

$$\big|\partial^\alpha_{x,t,\omega} J_a\big|_0 \leq C_\alpha \frac{\varepsilon}{t^2} \log t \tag{2.3.4}$$

avec de plus à l'extérieur

$$\big\|\partial^\alpha_{x,t,\omega} J_a\big\|_0 \leq C_\alpha \ \varepsilon^{q+1-\lambda\left(\frac{q-3}{2}\right)} . \tag{2.3.5}$$

2.4. Bilan des erreurs en périodes I et II. Les estimations indiquées ci-dessus permettent d'établir sans difficulté le théorème suivant.

THÉORÈME 2.4.1. *Pour tout A, $0 < A < A_0$, il existe $\varepsilon_A > 0$ tel que, pour tout $0 < \varepsilon \leq \varepsilon_A$ et $t \leq \dfrac{A^2}{\varepsilon^2}$, on ait*

(i) $\big|\partial^\alpha_{x,t,\omega}(u - u_a)\big|_0 \leq C_\alpha \ \varepsilon^{23/9} \, |\log \varepsilon|$.

(ii) *Pour tout p, il existe q tel que l'on ait de plus, à l'extérieur,*

$$\big\|\partial^\alpha_{x,t,\omega}(u - u_a)\big\|_0 \leq C_\alpha \ \varepsilon^p.$$

Rappelons que q apparaît en (2.2.6). On choisit dorénavant q en sorte que ii) ait lieu pour $p = 8$.

2.5. Description en période III. Dans toute la suite on posera $z = \frac{1}{r}$, $\tau = \varepsilon\sqrt{t}$, $\zeta = \varepsilon^2 \log t$. Pour une fonction $S(\sigma, \omega, z, \tau, \zeta)$ régulière, on notera $S' = \partial_\sigma S$, $S'' = \partial_\sigma^2 S$, etc., et $\dot{S} = \partial_\tau S$, $\ddot{S} = \partial_\tau^2 S$, etc.

2.5.1. L'hypothèse (ND) sur le minimum de $-g\partial_\sigma^2 R$ assure que la solution $u_a^{II,i}$ définie par (2.3.2) n'explose pas en période III. Nous prenons donc

$$u_a^{III,i} = u_a^{II,i} . \tag{2.5.1}$$

2.5.2. Notons, pour la fonction F de (2.3.1),

$$F_A(\sigma,\omega) = F\Big(\sigma,\omega,\frac{\varepsilon^2}{A^2}\Big(1+\frac{\sigma\varepsilon^2}{A^2}\Big)^{-1}, A, 2\varepsilon^2\log\frac{A}{\varepsilon}\Big), \tag{2.5.2}$$

et définissons (par abus) une nouvelle fonction $F(\sigma,\omega,\tau)$ par

$$\partial_\tau F - \frac{g}{2}(\partial_\sigma F)^2 = 0, \qquad F(\sigma,\omega,A) = F_A(\sigma,\omega). \tag{2.5.3}$$

D'autre part, en notant

$$E(F) = -\frac{1}{\tau^3}\Big(\frac{1}{4}+\partial_\omega^2\Big)F - \frac{\sigma g}{2\tau^2}F'F'' - \frac{\partial_\tau F}{4\tau^2} + \frac{\partial_\tau^2 F}{4\tau} + g^k_{ij}(\omega_k F' A_{ij} F' + \omega_i\omega_j F'' A_k F)$$

l'expression introduite au lemme 6.1.1 (dans laquelle les A_k, A_{ij} sont des opérateurs d'ordre 1 en ∂_ω, ∂_τ), définissons une fonction $G(\sigma,\omega,\tau)$ par

$$\partial_\tau G - g(\partial_\sigma F)\partial_\sigma G = \int_M^\sigma E(F)\,ds, \qquad G(\sigma,\omega,A) = 0. \tag{2.5.4}$$

Nous choisissons alors

$$u_a^{III,e} = \frac{\varepsilon}{r^{1/2}}(F+\varepsilon^2 G)(r-t,\omega,\tau). \tag{2.5.5}$$

2.5.3. Le raccord entre les périodes II et III se traduit par l'estimation

$$\big\|\partial_{x,\omega}^\alpha(\partial_t^+)^k(u-u_a)\big\|_0 \le C_{\alpha,k}\varepsilon^8 \text{ pour } \tau = A, \tag{2.5.6}$$

où ∂_t^+ désigne la dérivée à droite en $t = \dfrac{A^2}{\varepsilon^2}$.

2.5.4. Après recollement, on obtient pour J_a les estimations

$$\text{A l'intérieur, } \big|\partial_{x,t,\omega}^\alpha J_a\big|_0 \le C_\alpha\varepsilon^5\,|\log\varepsilon|\,, \tag{2.5.7}$$

$$\left\{\begin{array}{ll} & \text{A l'extérieur, sous l'hypothèse (ND)} \\ & \text{(ou (ND)'), pour un } \nu_1 > 0, \\ \text{(i)} & |J_a|_0 \le C\dfrac{\varepsilon^7}{(\tilde\tau-\tau)^{3-\nu_1}}, \\ \text{(ii)} & |\nabla J_a|_0 + |\partial_\omega J_a|_0 \le C\dfrac{\varepsilon^7}{(\tilde\tau-\tau)^{9/2-\nu_1}}, \\ \text{(iii)} & |\nabla^2 J_a|_0 + |\partial_\omega\nabla J_a|_0 + |\partial_\omega^2 J_a|_0 \le C\dfrac{\varepsilon^7}{(\tilde\tau-\tau)^{6-\nu_1}}. \end{array}\right. \tag{2.5.8}$$

On a posé ici $\tilde\tau = \tilde\tau(\varepsilon) = \tau_*(\varepsilon) - C\varepsilon^{\kappa/\kappa-1} - C\varepsilon^2\,|\log\varepsilon|$, pour un C assez grand (cf. lemme 6.4.1).

2.5.5. La solution approchée u_a possède en outre les propriétés suivantes:

$$\text{A l'intérieur } \big\|\partial_{x,t,\omega}^\alpha u_a\big\|_0 \le C_\alpha\varepsilon^2\ . \tag{2.5.9}$$

$$
(2.5.10)\qquad \left\{\begin{array}{ll} & \text{A l'extérieur,}\\ \text{(i)} & |u_a| + |\nabla_{x,t,\omega} u_a| \le C\varepsilon^2\\ \text{(ii)} & \text{Pour } k \ge 2,\\ & \displaystyle\sum_{|\alpha|\le k} \left|\partial^\alpha_{x,t,\omega} u_a\right| \le C_\alpha \frac{\varepsilon^2}{(\tilde\tau-\tau)^{\frac{3k}{2}-\frac12}}\\ \text{(iii)} & \displaystyle\left\|\partial^2_{ij} u_a(x,t) - \frac{\varepsilon}{r^{1/2}}\omega_i\omega_j F''(r-t,\omega,\tau)\right\|_0 \le C\frac{\varepsilon^4}{(\tilde\tau-\tau)^{5/2}}\ . \end{array}\right.
$$

3. La solution en période I. On écrit formellement $u = \varepsilon u_1 + \varepsilon^2 u_2 + \cdots$, et l'on choisit les u_j en sorte que

(i) $\Box u_1 = 0$, $u_1|_{t=0} = u_1^0$, $\partial_t u_1|_{t=0} = u_1^1$, puis, pour $j \ge 2$,

(ii) $\Box u_j + Q_j = 0$, $u_j|_{t=0} = u_j^0$, $\partial_t u_j|_{t=0} = u_j^1$, où $Q_p = \sum_{\ell+\ell'=p} g^k_{ij}\,\partial_k u_\ell\,\partial^2_{ij} u_{\ell'}$.

Pour alléger, on écrira souvent $g\partial u\partial^2 v$ au lieu de $\sum_{i,j,k} g^k_{ij}\,\partial_k u\,\partial^2_{ij} v$.

Nous allons maintenant choisir des approximations des u_j à l'extérieur et à l'intérieur.

3.1. Les approximations à l'extérieur. Elle reposent sur le lemme suivant.

LEMME 3.1.1. *Soit S une fonction C^∞ de ses arguments.*

(a) *Pour tous $\mu \in \mathbb{R}$ et $k \in \mathbb{N}$, on a l'identité*

$$
\begin{aligned}
(3.1.1)\quad & \frac{1}{t^\mu(\log t)^k}\Box\frac{t^\mu(\log t)^k}{r^{1/2}} S\Big(r-t,\omega,\frac1r,\tau,\zeta\Big)\\
&= \frac{-2S'}{t\sqrt r}\Big(\mu+\frac{k}{\log t}\Big) + \frac{QS}{r^{5/2}} + \frac{1}{t^2\sqrt r}\Big(\mu(\mu-1) + \frac{(2\mu-1)k}{\log t} + \frac{k(k-1)}{(\log t)^2}\Big) S\\
&\quad + \frac{\varepsilon}{\sqrt{rt}}\left[-\Big(\dot S' + \frac z4 \dot S\Big) + \frac{\dot S}{t}\Big(\mu+\frac{k}{\log t}\Big)\right]\\
&\quad + \frac{\varepsilon^2}{t\sqrt r}\left[\frac{\ddot S}{4} - 2\partial_\zeta S' - z\partial_\zeta S + \frac2t\partial_\zeta S\Big(\mu+\frac{k}{\log t}\Big)\right] + \frac{\varepsilon^3}{t^{3/2}\sqrt r}\partial_\zeta \dot S + \frac{\varepsilon^4}{t^2\sqrt r}\partial^2_\zeta S,
\end{aligned}
$$

où $QS = -\Big(\frac14 + \partial^2_\omega\Big) S + 2\partial_z S' - 2z\partial_z S' - z^2\partial^2_z S$.

(b) *Pour toute $S(r-t,\omega)$, tous $\mu \in \mathbb{R}$ et $k \in \mathbb{N}$ et tous $N \in \mathbb{N}$, $N' \in \mathbb{N}$, il existe des fonctions $\Sigma_\ell\Big(r-t,\omega,\frac1r\Big)$ $(0 \le \ell \le k)$ et $L_\ell\Big(r-t,\omega,\frac1r\Big)$ $(0 \le \ell \le k+1)$ telles que*

$$
\Box\frac{1}{r^{1/2}}\left\{\sum_{\ell\le k} t^{\mu+\frac32}(\log t)^\ell\Sigma_\ell + \sum_{\ell\le k+1}(\log t)^\ell L_\ell\right\} = t^\mu(\log t)^k S + O(z^N).
$$

De plus, $\Box\dfrac{L_\ell}{r^{1/2}} = O(z^{N'})$, et Σ_0 ne contient pas de terme en $\Big(\dfrac1r\Big)^{\mu+3/2}$.

Preuve. (a) Remarquons tout d'abord qu'à l'extérieur, on peut écrire $t = r-\sigma = r(1-\sigma z)$ et $r = t\big(1+\frac{\sigma}{t}\big)$, ce qui permet d'échanger asymptotiquement toute puissance de r et toute puissance de t. On a

$$
\begin{aligned}
\Box t^\mu(\log t)^k u = t^\mu(\log t)^k\Big\{&\Box u + \frac2t\Big(\mu+\frac{k}{\log t}\Big)\partial_t u\\
&+\frac{1}{t^2}\Big(\mu(\mu-1) + \frac{(2\mu-1)k}{\log t} + \frac{k(k-1)}{(\log t)^2}\Big) u\Big\}\ .
\end{aligned}
$$

D'autre part,

$$\Box \frac{S}{r^{1/2}} = \frac{1}{r^{1/2}} \Big\{ \partial_t^2 - \partial_r^2 - \frac{1}{r^2} A \Big\} S,$$

où $A = \frac{1}{4} + \partial_\omega^2$; comme

$$(\partial_t + \partial_r) S = -\frac{1}{r^2} \partial_z S + \frac{\varepsilon}{2\sqrt{t}} \dot{S} + \frac{\varepsilon^2}{t} \partial_\zeta S,$$

$$\begin{aligned}(\partial_t^2 - \partial_r^2) S = &- \frac{2}{r^3} \partial_z S - \frac{1}{r^2} \Big\{ -2 \partial_z S' + \frac{1}{r^2} \partial_z^2 S + \frac{\varepsilon}{2\sqrt{t}} \partial_z \dot{S} + \frac{\varepsilon^2}{t} \partial_{z\zeta}^2 S \Big\} \\ &- \frac{\varepsilon}{4t^{3/2}} \dot{S} + \frac{\varepsilon}{2\sqrt{t}} \Big\{ -2\dot{S}' + \frac{1}{r^2} \partial_z \dot{S} + \frac{\varepsilon}{2\sqrt{t}} \ddot{S} + \frac{\varepsilon^2}{t} \partial_\zeta \dot{S} \Big\} - \frac{\varepsilon^2}{t^2} \partial_\zeta S \\ &+ \frac{\varepsilon^2}{t^2} \Big\{ -2 \partial_\zeta S' + \frac{1}{r^2} \partial_{z\zeta}^2 S + \frac{\varepsilon}{2\sqrt{t}} \partial_\zeta \dot{S} + \frac{\varepsilon^2}{t} \partial_\zeta^2 S \Big\},\end{aligned}$$

et on trouve

$$\begin{aligned}\Box \frac{S}{r^{1/2}} = \frac{1}{r^{5/2}} QS + \frac{\varepsilon}{\sqrt{rt}} \Big(-\dot{S}' - \frac{z}{4} \dot{S} \Big) + \frac{\varepsilon^2}{t\sqrt{r}} \Big(\frac{\ddot{S}}{4} - 2 \partial_\zeta S' - z \partial_\zeta S \Big) \\ + \frac{\varepsilon^3}{t^{3/2}\sqrt{r}} \partial_\zeta \dot{S} + \frac{\varepsilon^4}{t^2 \sqrt{r}} \partial_\zeta^2 S \ .\end{aligned}$$

(b) A l'aide de la formule (3.1.1) (appliquée à des fonctions indépendantes de τ, ζ), on obtient, si $\mu \neq -\frac{3}{2}$,

$$\Box \frac{t^{\mu+\frac{3}{2}}}{r^{1/2}} \left\{ \sum_{0 \leq \ell \leq k} (\log t)^\ell \Sigma_\ell (r-t, \omega) \right\} = t^\mu (\log t)^k S + \sum_{\substack{\ell \geq 0 \\ 0 \leq \ell' \leq k}} t^{\mu-1-\ell} (\log t)^{\ell'} S_{\ell,\ell'} (r-t, \omega),$$

avec

$$\Sigma_k(\sigma, \omega) = \frac{-1}{2(\mu + \frac{3}{2})} \int_M^\sigma S(s, \omega)\, ds, \text{etc.}$$

Si $\mu = -\frac{3}{2}$, on obtient

$$\Box \frac{(\log t)^{k+1} L(r-t, \omega, \frac{1}{r})}{r^{1/2}} = t^\mu (\log t)^k S + \sum_{\substack{\ell \geq 0 \\ \ell' \leq k}} t^{\mu-1-\ell} (\log t)^{\ell'} S_{\ell,\ell'} (r-t, \omega) + O(z^N),$$

en choisissant

$$L(\sigma, \omega, 0) = \frac{-1}{2(k+1)} \int_M^\sigma S(s, \omega)\, ds,$$

et L solution, à un ordre convenable en z, de $QL = 0$.

En répétant cette procédure, on obtient, si $\mu + \frac{3}{2} \notin \mathbb{N}$,

$$\Box \frac{1}{r^{1/2}} \Big(\sum_{\substack{\ell' \leq k \\ 0 \leq \ell \leq N}} t^{\mu+\frac{3}{2}-\ell} (\log t)^{\ell'} \Sigma_{\ell,\ell'} (r-t, \omega) \Big) = t^\mu (\log t)^k S + O(z^{N+1-\mu}).$$

Si $\mu+\frac{3}{2}=q\in\mathbb{N}$, on trouve

$$\Box\frac{1}{r^{1/2}}\Big\{\sum_{\substack{\ell'\le k\\ 0\le\ell\le N}} t^{\mu+\frac{3}{2}-\ell}(\log t)^{\ell'}\Sigma_{\ell,\ell'}(r-t,\omega) + \sum_{1\le\ell'\le k+1}(\log t)^{\ell'}L_{\ell'}\Big(r-t,\omega,\frac{1}{r}\Big)\Big\} = t^{\mu}(\log t)^k S + O(z^{N+1-\mu}).$$

De plus, $\Sigma_{q,0}=0$. Dans tous les cas, on réécrit le premier terme de la solution obtenue sous la forme

$$\sum_{\ell\le k}\frac{t^{\mu+\frac{3}{2}}}{r^{1/2}}(\log t)^{\ell}\Sigma_{\ell}\Big(r-t,\omega,\frac{1}{r}\Big)\ ,$$

Σ_0 vérifiant la condition du lemme. □

Nous sommes alors en mesure de prouver la proposition suivante.

PROPOSITION 3.1.1. *Pour tous $N,N'\in\mathbb{N}$, le terme u_k peut s'écrire*

$$u_k=\frac{1}{r^{1/2}}\Big\{\sum_{\substack{\ell\ge 0\\ 2\ell\le k-1}}(\log t)^{\ell}L_k^{\ell}\Big(r-t,\omega,\frac{1}{r}\Big) + \sum_{\substack{\ell\ge 1\\ \ell+2\ell'\le k-1}} t^{\ell/2}(\log t)^{\ell'}R_k^{\ell,\ell'}\Big(r-t,\omega,\frac{1}{r}\Big)\Big\}+r_k,$$

où

$$r_k=O(z^N),\quad \textit{et}\quad \Box\frac{L_k^{\ell}}{r^{1/2}}=O(z^{N'}).$$

De plus, si $\ell=2p$ $(p\ge 1)$, $R_k^{\ell,0}$ ne contient pas de terme en $\big(\frac{1}{r_k}\big)^p$.

Enfin, si l'on pose $U_k=\sum_{\ell=1}\varepsilon^{\ell}u_{\ell}$, on a

$$\mathcal{F}_k\equiv\Box U_k+g\,\partial U_k\,\partial^2U_k=\frac{\varepsilon^2}{r}\sum_{\substack{h\ge k-1\\ m+2p\le h}}\varepsilon^{h-m-2p}\tau^m\zeta^pF_{m,p}(\sigma,\omega,z),$$

avec la majoration $\mathcal{F}_k=O\big(\varepsilon^{k+1}t^{(k-3)/2}\big)$.

Preuve. (a) Pour $k=1$, on a $u_1=\frac{1}{r^{1/2}}R_0^{0,0}\big(r-t,\omega,\frac{1}{r}\big)$ (voir [5]). Procédons alors par récurrence, en supposant correcte la forme de u_j pour $j\le k-1$. Comme toute dérivée (en x ou t) de u_j a encore la même forme (en oubliant cette fois que $\frac{L_k^{\ell}}{r^{1/2}}$ est dans le noyau de $\Box$), on obtient

$$Q_k=\sum\frac{1}{r}t^{\ell/2}(\log t)^{\ell'}t^{q/2}(\log t)^{q'}R_j^{\ell,\ell'}R_{j'}^{q,q'},$$

où la somme $\sum$ est étendue à tous les indices ℓ,ℓ',q,q',j,j' tels que $\ell\ge 0$, $q\ge 0$, $\ell+2\ell'\le j-1$, $q+2q'\le j'-1$, $j+j'=k$.

En développant les fonctions R, on trouve formellement

$$Q_k=\sum_{p\ge 0,\,m\ge 0}t^{(m-2\ell)/2}(\log t)^pS_{m,p}(r-t,\omega),\quad \ell\ge 1,\quad m+2p\le k-2.$$

Le lemme 3.1.1 permet alors de "résoudre" $\Box v_k + Q_k = 0$ sous la forme

$$v_k = \frac{1}{r^{1/2}} \Big\{ \sum_{\substack{m+2p \le k-1 \\ m \ge 1}} t^{m/2} (\log t)^p \Sigma_{m,p}\Big(r-t, \omega, \frac{1}{r}\Big) + \sum_{\substack{2p \le k-1 \\ p \ge 1}} (\log t)^p L_p \Big\}.$$

En effet, les termes en $t^{-3/2}(\log t)^p$ dans Q_k ne peuvent apparaître que pour $m \ge 1$, donc pour $2p \le k-3$; le terme correspondant dans v_k est alors en $(\log t)^{p+1}$, où $2(p+1) \le k-1$. D'autre part, pour $m-2\ell \neq -3$, le terme de v_k correspondant au terme $t^{(m-2\ell)/2}(\log t)^p$ de Q_k est en $t^{(m+3-2\ell)/2}(\log t)^p$; on écrit $t^{(m+3-2\ell)/2} = t^{(m+1)/2} \frac{1}{t^{\ell-1}}$, et $(m+1)+2p \le k-1$ comme annoncé.

(b) Au point (a), on a obtenu v_k telle que $\Box v_k + Q_k$ soit de l'ordre d'une puissance de z arbitraire. On a donc $\Box(u_k - v_k)$ de décroissance arbitraire à l'extérieur: par vitesse finie de propagation $u_k - v_k$ coïncide à l'extérieur avec la solution w de $\Box w = f$, où f est de décroissance arbitraire. Par une adaptation immédiate de [1], on obtient que w diffère d'une solution libre par une fonction de décroissance arbitraire: $u_k - v_k = (L_k^0(r-t, \omega, \frac{1}{r}))/r^{1/2} + r_k$. La forme de u_k est donc celle annoncée.

(c) Comme $\Box U_k + g\partial U_k \partial^2 U_k = \sum_{k+1 \le \ell+\ell' \le 2k} \varepsilon^{\ell+\ell'} g\partial u_\ell \partial^2 u_{\ell'}$, et $|\partial^\alpha u_\ell| \le C \dfrac{t^{(\ell-1)/2}}{r^{1/2}}$, on a $\Box U_k + g\partial U_k \partial^2 U_k = O\Big(\varepsilon^{k+1} t^{\frac{k-3}{2}}\Big)$.

On peut aussi écrire

$$\varepsilon^k u_k = \frac{\varepsilon}{r^{1/2}} \sum \varepsilon^{k-1-\ell-2\ell'} \tau^\ell \zeta^{\ell'} R_k^{\ell,\ell'}(\sigma, \omega, z) \ ,$$

et

$$\mathcal{F}_k = \frac{\varepsilon^2}{r} \sum_{\substack{m \ge 0 \\ p \ge 0 \\ h \ge k-1 \\ m+2p \le h}} \varepsilon^{h-m-2p} \tau^m \zeta^p F_{m,p}(\sigma, \omega, z). \qquad \square$$

3.2. Les approximations à l'intérieur. Ici, on modifie le point de vue du §3.1, qui ne peut donner de résultats qu'à l'extérieur (voir §2.1) ; nous nous contentons de choisir pour les termes u_k^i des approximations assez grossières des u_k, en portant une attention spéciale au comportement symbolique en σ et à l'approximation des solutions d'équations d'ondes non homogènes.

Les développements de cette section sont très proches de la partie correspondante de [1], dans le cas irrotationnel $\omega \equiv 0$.

Le fait de travailler avec des données *générales* (c'est-à-dire non nécessairement invariantes par rotation) n'apporte aucune modification non triviale ; en particulier, le paragraphe 7 de [1] consacré au comportement des solutions de l'équation des ondes non homogène s'étend sans difficulté au cas général, avec estimation des dérivées ∂_ω^α (outre les dérivées en $\partial_{x,t}^\alpha$). Néanmoins, il est nécessaire ici d'augmenter la précision de la solution approchée, ce qui entraîne quelques complications techniques.

Nous nous contentons ici d'insister sur les aspects nouveaux de la construction, renvoyant le lecteur à [1] pour les généralités. Nous noterons génériquement S_m une fonction $S(r-t, \omega)$ qui se comporte comme un symbole d'ordre m en la première variable $\sigma = r - t$.

3.2.1 On choisit pour u_1^i la solution de $\Box u_1 = 0$, avec des données $u_1|_{t=0} = u_1^0$, $\partial_t u_1|_{t=0} = u_1^1$, c'est-à-dire $u_1^i \equiv u_1$.

Le comportement $u_1 \sim (R(r-t, \omega))/r^{1/2}$ est établi dans [4], [5].

3.2.2 On choisit pour u_2^i la solution de $\Box u_2 + g_{ij}^k \partial_k u_1 \partial_{ij}^2 u_1 = 0$, avec des données $u_2|_{t=0} = u_2^0$, $\partial_t u_2|_{t=0} = u_2^1$, c'est-à-dire $u_2^i \equiv u_2$.

Comme en [1], on introduit une troncature $\chi = \chi\Big(r/(1+t)\Big)$, où $\chi \in C^\infty(\mathbb{R})$, $\chi(s) = 0$ pour $s \leq \frac{1}{2}$, $\chi(s) = 1$ pour $s \geq \frac{2}{3}$.

Le lemme suivant précise le comportement asymptotique de u_2.

LEMME 3.2.1. *On a $u_2 = \overline{u}_2 + z$, avec $\overline{u}_2 = (1-\theta(t))\frac{g}{2}\sqrt{t/r}\,\chi R'^2$, (rappelons que $\theta \in C^\infty(\mathbb{R})$ vaut 1 près de $t=0$) et z vérifiant les propriétés:*

Il existe une fonction $L(\sigma,\omega)$ de classe C^∞, supportée dans $\sigma \leq M$, avec $\left|\partial_\omega^k L\right| \leq C_k(1+|\sigma|)^{1/2}$, et pour $\ell \geq 1$,

$$\int \left|\partial_\omega^k \partial_\sigma^\ell L\right|^2 (\sigma,\omega)\,d\sigma \leq C_{k,\ell}, \qquad \left|\partial_\omega^k \partial_\sigma^\ell L\right| \leq C_{k,\ell}\ ,$$

telle que

(i) *Pour $r \leq Ct$, $(C<1)$, $|\alpha| \geq 1$,*

$$\left|\partial_{x,t}^\alpha \partial_\omega^\beta z\right| \leq \frac{C}{(1+t)^{1+\frac{3}{4}\inf(|\alpha|,4)}}.$$

(ii) *Pour $r \geq Ct$, $(C<1)$, $|\alpha| \geq 1$,*

$$\left|\partial_{x,t}^\alpha \partial_\omega^\beta \Big(z - \frac{L(r-t,\omega)}{r^{1/2}}\Big)\right| \leq C_{\alpha,\beta}\Big\{\frac{S_{-1/2} + h_t(r-t,\omega)}{1+t} + \frac{S_{1/2}}{(1+t)^{3/2}}\Big\},$$

où $\int |h_t(\sigma,\omega)|^2\,d\sigma \leq 1$.

Preuve. (a) Il faut d'abord établir une forme précise du terme quadratique

$$Q_2 = g_{ij}^k \partial_k u_1 \partial_{ij}^2 u_1\ .$$

D'après [5], on sait que $u_1 = \frac{1}{r^{1/2}} F\Big(r-t,\omega,\frac{1}{r}\Big)$, où

$$\left|\partial_\omega^\alpha \partial_z^\beta \partial_\sigma^\gamma F(\sigma,\omega,z)\right| \leq C(1+|\sigma|)^{-\frac{1}{2}+|\beta|-|\gamma|}\ ,$$

et par ailleurs, $\left|\partial_{x,t}^\alpha u_1\right| \leq C_\alpha/(1+t)^{1+|\alpha|}$ pour $r \leq Ct$, $(C<1)$.

- Pour $r \leq Ct$, $(C<1)$, on a donc $Q_2 = O\Big(1/(1+t)^5\Big)$.
- Pour $r \geq Ct$, on peut écrire

$$u_1 = \frac{S_{-1/2}}{r^{1/2}} + \frac{S_{1/2}}{r^{3/2}} + \cdots + \frac{S_{\frac{k}{2}-1}}{r^{k/2}} + \cdots$$

$$\nabla u_1 = \frac{S_{-3/2}}{r^{1/2}} + \frac{S_{-1/2}}{r^{3/2}} + \cdots + \frac{S_{\frac{k}{2}-2}}{r^{k/2}} + \cdots, \qquad \nabla^2 u_1 = \frac{S_{-5/2}}{r^{1/2}} + \cdots + \frac{S_{\frac{k}{2}-3}}{r^{k/2}} + \cdots$$

On obtient donc

$$Q_2 = \frac{S_{-4}}{r} + \frac{S_{-3}}{r^2} + \frac{S_{-2}}{r^3} + \frac{S_{-1}}{r^4} + \frac{S_0}{r^5} + \cdots,$$

le premier terme valant en fait $g(\omega)\dfrac{R'R''}{r}$.

(b) On va maintenant relever les termes de Q_2. Remarquons que si

$$L \in S_m, \Box\chi\frac{L}{r^{1/2}} = \chi\frac{S_m}{r^{5/2}} + O\Big(\frac{1}{t^{5/2-m}}\Big),$$

donc

$$\Box\Big(\chi\frac{L}{r^{1/2}}t^\mu\Big) = t^{\mu-2}\chi\frac{S_m}{r^{1/2}} - 2\mu\chi\frac{L't^{\mu-1}}{r^{1/2}} + O\Big(\frac{1}{t^{\frac{5}{2}-m-\mu}}\Big).$$

Comme

$$t^\alpha = r^\alpha + \sum_{k\geq 1}\frac{S_{+k}}{r^{k-\alpha}},$$

on a finalement

$$\text{(3.2.1)}\quad \Box\Big(\chi\frac{L}{r^{1/2}}t^\mu\Big) = -2\mu\chi\frac{L'}{r^{3/2-\mu}} + \chi\Big(\frac{S_m}{r^{5/2-\mu}} + \frac{S_{m+1}}{r^{3/2-\mu}} + \cdots\Big) + O\Big(\frac{1}{t^{5/2-m-\mu}}\Big).$$

On en déduit, en commençant avec

$$\mu = \frac{1}{2},\quad m = -3,\quad \Box\overline{u}_2 + Q_2 = \frac{S_{-3}}{r^2} + \frac{S_{-2}}{r^3} + \cdots + O\Big(\frac{1}{t^5}\Big);$$

en relevant les termes successifs, il vient

$$\Box\Big(\overline{u}_2 + \chi\Big(\frac{S_{-2}}{r} + \frac{S_{-1}}{r^2} + \frac{S_0}{r^3}\Big)\Big) + Q_2 = O\Big(\frac{1}{t^5}\Big)\ .$$

(c) La solution de $\Box v = O\Big(\dfrac{1}{t^5}\Big)$ peut être étudiée à l'aide du paragraphe 7 de [1]: on a dans ce cas

$$\Big|O\Big(\frac{1}{t^5}\Big)\Big|_0 \leq \frac{C}{t^4},$$

donc $\lambda = 3$. Les estimations du lemme, qui sont trivialement vraies avec $L \equiv 0$ pour la partie explicite

$$\frac{S_{-2}}{r} + \frac{S_{-1}}{r^2} + \frac{S_0}{r^3}$$

de z, résultent alors du lemme 7 de [1], points b et c. □

3.2.3. Nous posons

$$\overline{u}_3 = (1-\theta(t))\chi\frac{t}{\sqrt{r}}\frac{g^2}{6}(R'^3)', \qquad \bar{\bar{u}}_3 = (1-\theta(t))\chi g\sqrt{\frac{t}{r}}R'L',$$

et nous choisissons $u_3^i = \overline{u}_3 + \bar{\bar{u}}_3$.

Les raisons de ce choix résultent d'une analyse du terme d'interaction cubique $Q_3 = g_{ij}^k\,\partial_k u_1\,\partial_{ij}^2 u_2 + g_{ij}^k\,\partial_k u_2\,\partial_{ij}^2 u_1$, qui est résumée dans le lemme suivant.

LEMME 3.2.2. *On a*

$$Q_3 = \chi\frac{g^2}{3}(R'^3)''\frac{\sqrt{t}}{r} + \chi\frac{g}{r}(R'L')' + W_1$$

et $\Box u_3 + Q_3 = W_2$, *avec*

$$\left|\partial^\alpha_{x,t,\omega} W_1\right|_0 + \left|\partial^\alpha_{x,t,\omega} W_2\right|_0 \leq \frac{C_\alpha}{1+t} \ .$$

Preuve. (a) On a

$$\begin{aligned} Q_3 = {} & g^k_{ij}\,\partial_k\Big(\chi\frac{R}{r^{1/2}}\Big)\partial^2_{ij}\Big(\overline{u}_2 + \chi\frac{L}{r^{1/2}}\Big) + g^k_{ij}\,\partial_k\Big(\overline{u}_2 + \chi\frac{L}{r^{1/2}}\Big)\partial^2_{ij}\Big(\chi\frac{R}{r^{1/2}}\Big) \\ & + g^k_{ij}\,\partial_k\Big(u_1 - \chi\frac{R}{r^{1/2}}\Big)\partial^2_{ij}u_2 + g^k_{ij}\,\partial_k u_2\,\partial^2_{ij}\Big(u_1 - \chi\frac{R}{r^{1/2}}\Big) \\ & + g^k_{ij}\,\partial_k\Big(\chi\frac{R}{r^{1/2}}\Big)\partial^2_{ij}\Big(z - \chi\frac{L}{r^{1/2}}\Big) + g^k_{ij}\,\partial_k\Big(z - \chi\frac{L}{r^{1/2}}\Big)\partial^2_{ij}\Big(\chi\frac{R}{r^{1/2}}\Big) \\ \equiv {} & Q^0_3 + Q^1_3 + Q^2_3. \end{aligned}$$

• Comme

$$|\partial^\alpha_{\overline{u}_2}|_0 \leq C(1+t)^{1/2}, \left|\partial^\alpha\Big(u_1 - \chi\frac{R}{r^{1/2}}\Big)\right| \leq \frac{C}{(1+t)^{3/2}}$$

si $|\alpha| \geq 1$, on a facilement

$$|\partial^\alpha Q^1_3|_0 \leq \frac{C}{1+t}.$$

• L'estimation de Q^2_3 est plus délicate, et fait appel au lemme 3.2 1: on obtient

$$|\partial^\alpha Q^2_3|_0 \leq \frac{C}{1+t}.$$

(b) Dans Q^0_3, les termes obtenus en faisant porter les dérivées uniquement sur la variable $r-t$ des symboles valent

$$\chi^2\frac{g^2}{3}(R'^3)''\frac{\sqrt{t}}{r} + \chi^2\frac{g}{r}(R'L')'.$$

Les différences avec les termes annoncés sont de la forme

$$(\chi - \chi^2)\frac{1}{t^{1/2}}S_{-13/2} + (\chi - \chi^2)\frac{1}{t}S_{-3/2} = O\Big(\frac{1}{t^{5/2}}\Big),$$

leurs normes L^2 sont bornées par $C/(1+t)^{3/2}$.

Tous les autres termes faisant intervenir R et $\overline{u}_2$ sont de la forme $\dfrac{S_m}{t^{3/2}}$, avec $m \leq -\dfrac{7}{2}$ ou sont majorés par $\dfrac{1}{t^5}$; tous les autres termes faisant intervenir R et L sont majorés par $\dfrac{S_{-3/2}}{t^2}$, ou sont $O\Big(\frac{1}{t^{9/2}}\Big)$. Tous ces termes sont bornés en norme L^2 par $\dfrac{C}{1+t}$. □

3.2.4. L'analyse des termes d'interactions $Q_\ell = \sum_{\ell'+\ell''=\ell} g^k_{ij}\,\partial_k u_{\ell'}\,\partial^2_{ij}u_{\ell''}$ devient vite fastidieuse pour $\ell \geq 4$. Il importe donc de systématiser un peu la construction.

• La forme de la solution en zone II ($t \geq \varepsilon^{-\lambda}$) et l'exigence $\|\nabla\dot{u}\|_0 = o(\varepsilon^{7/2})$ rendent nécessaire de choisir $\lambda > \dfrac{3}{2}$. Les termes de Q_ℓ qui sont négligeables sont ceux dont l'intégrale de la norme L^2 jusqu'au temps $\varepsilon^{-\lambda}$ est $o(\varepsilon^{5/2-\ell})$.

• D'autre part, en supposant le terme principal de u_ℓ de la forme $\overline{u}_\ell = \chi a_\ell^1(r-t,\omega)t^{\frac{1}{2}(\ell-2)}$, on trouve que le terme principal de Q_ℓ vaut $\chi t^{\frac{1}{2}(\ell-4)} g \sum (a_{\ell'}^1)'(a_{\ell''}^1)''$. On en déduit par récurrence que $\overline{u}_\ell$ a bien la forme voulue (grâce à la formule (3.2.1)).

• Des deux points précédents on déduit que Q_ℓ lui-même est négligeable si $t^{1/2(\ell-4)+1/2+1}|_{t=\varepsilon^{-\lambda}} = o(\varepsilon^{5/2-\ell})$, i.e. $-\lambda/2(\ell-1) > \frac{5}{2} - \ell$, soit $\ell > 7$ pour λ proche de $\frac{3}{2}$.

• De même, les termes non principaux de Q_ℓ, qui sont de la forme $S_m \times t^{1/2(\ell-4)-1/2}$, sont négligeables si $-\frac{\lambda}{2}(\ell-2) > \frac{5}{2} - \ell$, soit $\ell > 4$ pour λ proche de $\frac{3}{2}$.

Nous devons donc encore analyser en détail Q_4, exactement comme nous l'avons fait pour Q_3 au lemme 3.1. (c).

En notant $\overline{u}_4 = \chi \frac{g^3}{6} \frac{t^{3/2}}{r^{1/2}} (3R'^2 R''^2 + R'^3 R''')$, $\bar{\bar{u}}_4 = \chi \frac{g^2}{2} \frac{t}{r^{1/2}} (2R' R'' L' + R'^2 L'')$, on choisit $u_4^i = \overline{u}_4 + \bar{\bar{u}}_4$.

LEMME 3.2.3. *On a*

$$\Box u_4 + Q_4 = W_3 \ , \ \textit{avec} \ \left|\partial_{x,t,\omega}^\alpha W_3\right|_0 \leq \frac{C_\alpha}{(1+t)^{1/2}} \ .$$

Preuve. (a) On a

$$\begin{aligned}
Q_4 &= g_{ij}^k \partial_k u_2 \partial_{ij}^2 u_2 + g_{ij}^k \partial_k u_1 \partial_{ij}^2 u_3 + g_{ij}^k \partial_k u_3 \partial_{ij}^2 u_1 \\
&= \Big[g_{ij}^k \partial_k \overline{u}_2 \partial_{ij}^2 \overline{u}_2 + g_{ij}^k \partial_k \overline{u}_2 \partial_{ij}^2 \Big(\chi \frac{L}{r^{1/2}}\Big) \\
&+ g_{ij}^k \partial_k u_2 \partial_{ij}^2 \Big(z - \chi \frac{L}{r^{1/2}}\Big) + g_{ij}^k \partial_k \Big(\chi \frac{L}{r^{1/2}}\Big) \partial_{ij}^2 \overline{u}_2 + g_{ij}^k \partial_k \Big(z - \chi \frac{L}{r^{1/2}}\Big) \partial_{ij}^2 \overline{u}_2 \\
&+ g_{ij}^k \partial_k \Big(\chi \frac{L}{r^{1/2}}\Big) \partial_{ij}^2 \Big(\chi \frac{L}{r^{1/2}}\Big) + g_{ij}^k \partial_k \Big(z - \chi \frac{L}{r^{1/2}}\Big) \partial_{ij}^2 \Big(\chi \frac{L}{r^{1/2}}\Big) \Big] \\
&+ \Big[g_{ij}^k \partial_k \Big(\chi \frac{R}{r^{1/2}}\Big) \Big(\partial_{ij}^2 \overline{u}_3 + \partial_{ij}^2 \bar{\bar{u}}_3\Big) + g_{ij}^k \partial_k (\overline{u}_3 + \bar{\bar{u}}_3) \partial_{ij}^2 \Big(\chi \frac{R}{r^{1/2}}\Big) \\
&+ g_{ij}^k \partial_k \Big(u_1 - \chi \frac{R}{r^{1/2}}\Big) \partial_{ij}^2 u_3 + g_{ij}^k \partial_k u_3 \partial_{ij}^2 \Big(u_1 - \chi \frac{R}{r^{1/2}}\Big) \Big] = Q_4^0 + Q_4^1 \ .
\end{aligned}$$

Comme

$$\left|\partial^\alpha \Big(\partial_k \overline{u}_2 \partial_{ij}^2 \overline{u}_2 - \frac{g^2}{4} \frac{t}{r} \chi^2 \omega_k \omega_i \omega_j (R'^2)'(R'^2)''\Big)\right|_0 \leq \frac{C}{(1+t)^{1/2}} \ ,$$

$$\left|\partial^\alpha \Big(\partial_k \overline{u}_2 \partial_{ij}^2 \Big(\chi \frac{L}{r^{1/2}}\Big) - \chi^2 \frac{g}{2} \frac{\sqrt{t}}{r} \omega_k \omega_i \omega_j (R'^2)' L''\Big)\right|_0 \leq \frac{C}{1+t} \ ,$$

et que les autres types de termes Q_4^0 ont leur norme L^2 bornée par $\frac{C}{(1+t)^{1/2}}$, il vient

$$\left|\partial^\alpha \Big[Q_4^0 - \chi \Big\{ \frac{g^3}{2} (R'^2 R''^2)' + \frac{g^2}{r^{1/2}} (R' R'' L')' \Big\}\Big]\right|_0 \leq \frac{C}{(1+t)^{1/2}} \ .$$

De même

$$\left|\partial^\alpha \Big(\partial_k \Big(\chi \frac{R}{r^{1/2}}\Big) \partial_{ij}^2 \overline{u}_3 - \chi^2 \frac{g^2}{6} R' (R'^3)''' \omega_k \omega_i \omega_j\Big)\right|_0 \leq \frac{C}{(1+t)^{1/2}} \ ,$$

$$\left|\partial^\alpha\left(\partial_k\left(\chi\frac{R}{r^{1/2}}\right)\partial_{ij}^2\bar{\bar{u}}_3-\chi^2\frac{g}{r^{1/2}}R'(R'L')''\omega_k\omega_i\omega_j\right)\right|_0\leq\frac{C}{1+t}\ ,$$

et tous les autres termes de Q_4^1 ont leur norme L^2 bornée par $\dfrac{C}{(1+t)^{1/2}}$, d'où

$$\left|\partial^\alpha\left[Q_4^1-\chi\left\{\frac{g^3}{6}(R'(R'^3)'')'+\frac{g^2}{r^{1/2}}(R'(R'L')')'\right\}\right]\right|_0\leq\frac{C}{(1+t)^{1/2}}\ .$$

(b) Par application de (3.2.1), on voit que les termes non principaux de $\Box u_4$ satisfont l'estimation du lemme 3.2.3. □

3.2.5. Les termes $\bar{u}_5$, $\bar{u}_6$ et $\bar{u}_7$ sont de la forme indiquée en §3.2.4, et

$$|\partial^\alpha(\Box\bar{u}_\ell+Q_\ell)|_0\leq C(1+t)^{\frac{1}{2}(\ell-4)}\ .$$

Nous pouvons résumer toute la construction du paragraphe 3.2 dans la proposition suivante.

PROPOSITION 3.2.1. *Posons* $u_a=u_a^I=\varepsilon u_1^i+\varepsilon^2u_2^i+\cdots+\varepsilon^7u_7^i$, *où les* u_j^i *ont été construits aux* §§3.2.1–3.2.5. *Alors, pour* $t\leq\dfrac{\text{cte}}{\varepsilon^2}$,

$$\left|\partial_{x,t,\omega}^\alpha J_a\right|_0\leq C_\alpha\left\{\frac{\varepsilon^3}{1+t}+\varepsilon^5(1+t)^{1/2}+\varepsilon^8(1+t)^{5/2}\right\}.$$

3.3. Recollement et estimation de l'erreur. On pose finalement, avec $k\geq 7$ à choisir,

$$u_a^I=(1-\theta(r-t+C_0-1))u_a^e+\theta(r-t+C_0-1)u_a^i,$$

où $u_a^i=\varepsilon u_1^i+\ldots+\varepsilon^7u_7^i$, et $u_a^e=U_k$.

On a alors l'estimation suivante de l'erreur $J_a=J_a^I=\Box u_a^I+g\partial u_a^I\partial^2u_a^I$.

PROPOSITION 3.3.1. *Pour tout* k *et* $t\leq\dfrac{\text{cte}}{\varepsilon^2}$, *on peut choisir* N *et* N' *en sorte que*

$$\left|\partial_{x,t,\omega}^\alpha J_a\right|_0\leq C_\alpha\left\{\frac{\varepsilon^3}{1+t}+\varepsilon^5(1+t)^{1/2}+\varepsilon^8(1+t)^{5/2}\right\}|\log t|^{k-1},$$

et de plus, pour $r-t\geq -C_0+1$,

$$\left\|\partial_{x,t,\omega}^\alpha J_a\right\|_0\leq C_\alpha\varepsilon^{k+1}t^{\frac{k-3}{2}}.$$

Preuve. L'estimation est celle de la proposition 3.1.1 à l'extérieur, celle de la proposition 3.2.1 à l'intérieur.

Remarquons que pour $3\leq j\leq 7$, la différence $u_j-u_j^i$ est formée de termes contenant une puissance non nulle de z ou de $(\log t)$, ou de termes indépendants de z et $(\log t)$ de la forme $S(r-t,\omega)/r^{1/2}t^{\ell/2}$, avec $0\leq\ell\leq j-3$. Donc

$$u^e-u^i=\sum_{3\leq j\leq 7}\frac{\varepsilon^j}{r^{1/2}}\left\{\sum_{\substack{\ell\geq 1\\ \ell\leq 2j-1}}(\log t)^\ell L_j^\ell+\sum_{\substack{\ell\geq 1\\ \ell'\geq 1\\ \ell+2\ell'\leq j-1}}t^{\ell/2}(\log t)^{\ell'}R_k^{\ell,\ell'}+\sum_{\ell\leq j-3}t^{\ell/2}R_k^{\ell,0}\right\}$$
$$+\sum_{8\leq j\leq k}\varepsilon^iu_j.$$

On en déduit

$$\|\partial^\alpha(u^e-u^i)\|_0 \le C\,|\log t|^{k-1}\left\{\frac{\varepsilon^3}{(1+t)^{1/2}}+\varepsilon^8 t^3\right\},$$

et

$$\|\partial^\alpha(\partial_t+\partial_r)(u^e-u^i)\|_0 \le C\,|\log t|^{k-1}\left\{\frac{\varepsilon^3}{(1+t)^{3/2}}+\varepsilon^8 t^2\right\}.$$

On écrit alors, avec $\delta = u^i - u^e$,

$$\Box(\theta u^i+(1-\theta)u^e) = \theta\Box u^i+(1-\theta)\Box u^e+[\Box,\theta]\delta\ ,$$

$$\begin{aligned} g\partial(\theta u^i+(1-\theta)u^e)\partial^2(\theta u^i+(1-\theta)u^e) &= g\partial(u^e+\theta\delta)\partial^2(u^e+\theta\delta)\\ &= g\,\partial u^e\partial^2 u^e+\theta\{g\,\partial u^e\partial^2\delta+g\,\partial\delta\partial^2 u^e\}+\theta^2 g\,\partial\delta\partial^2\delta+g[\partial,\theta]\delta\partial^2(\theta\delta)\\ &\quad+g\,\partial\delta[\partial^2,\theta]\delta+g\,\partial u^e[\partial^2,\theta]\delta+g[\partial,\theta]\delta\partial^2 u^e\\ &= (1-\theta)g\,\partial u^e\partial^2 u^e+\theta g\,\partial u^i\partial^2 u^i+\theta(\theta-1)g\,\partial\delta\partial^2\delta+\text{crochets}, \end{aligned}$$

en sorte que

$$J_a = \theta J_a^i+(1-\theta)J_a^e+\text{crochets}.$$

Comme $[\Box,\theta] = -2\theta'(\partial_t+\partial_r)-\dfrac{\theta'}{r}$, on a

$$\|[\Box,\theta]\delta\|_0 \le C\,|\log t|^{k-1}\left\{\frac{\varepsilon^3}{(1+t)^{3/2}}+\varepsilon^8 t^2\right\};$$

d'autre part, les termes quadratiques contenant des crochets sont bornés par $\dfrac{C\varepsilon}{r^{1/2}}\,\|\partial^\alpha\delta\|_0$ qui est encore inférieur, d'où l'estimation. □

4. La solution en période *II*. Soit dorénavant fixé A, $0<A<A_0$. Il s'agit maintenant de prolonger, pour $0\le\tau\le A$, la solution construite en période I.

4.1. La construction à l'extérieur. Remarquons que

$$u_a^e = U_k = \frac{\varepsilon}{r^{1/2}}\left\{\sum\varepsilon^{k-1-2\ell}\zeta^\ell L_k^\ell+\sum\varepsilon^{k-1-\ell-2\ell'}\tau^{\ell/2}\zeta^{\ell'}R_k^{\ell,\ell'}\right\}+\cdots$$

avec les propriétés supplémentaires indiquées à la proposition 3.1.

Cela conduit à chercher une solution approchée de l'équation sous la forme $u = \dfrac{\varepsilon}{r^{1/2}}F(\sigma,\omega,z,\tau,\zeta)$, avec $\sigma = r-t$, $z=\dfrac{1}{r}$, $\tau=\varepsilon\sqrt{t}$, $\zeta=\varepsilon^2\log t$, F dépendant elle même de ε (nous ne l'indiquons jamais) ; F devra en outre s'écrire $F = F_0+\tau G$, $F_0(\sigma,\omega,z,\zeta) = F(\sigma,\omega,z,0,\zeta)$ vérifiant, identiquement en ζ, l'équation

$$Q F_0 \equiv \left[-\left(\frac14+\partial_\omega^2\right)+2(1-z)\partial_{z\sigma}^2-z^2\partial_z^2\right]F_0 = 0 \quad \text{(cf. (3.1.1))}.$$

4.1.1 Les variables étant surabondantes, nous indiquons dans le lemme suivant le choix opéré.

LEMME 4.1.1. *Pour toute fonction F régulière, en posant*

$$u = \frac{\varepsilon}{r^{1/2}}F\Big(r-t,\omega,\frac1r,\varepsilon\sqrt{t},\varepsilon^2\log t\Big),$$

on a

$$\partial_i u = \frac{\varepsilon}{r^{1/2}}(\omega_i F' + z A_i F),\ \partial_{ij}^2 u = \frac{\varepsilon}{r^{1/2}}(\omega_i\omega_j F'' + z A_{ij} F)\ ,$$

où A_i *et* A_{ij} *sont des opérateurs différentiels en* $(\sigma, \omega, z, \tau, \zeta)$. *De plus,*

$$\Box u + g\partial_u \partial^2 u = \frac{\varepsilon}{r^{5/2}} Q F_0 + \frac{\varepsilon^2}{\sqrt{rt}}\Big\{ z(1-\sigma z) Q G - \Big(\dot F' + \frac{z}{4}\dot F\Big)$$
$$+\frac{\tau z}{1-\sigma z}\Big(\frac{\ddot F}{4} - 2\partial_\zeta F' - z\partial_\zeta F\Big) + \frac{\tau^2 z^2}{(1-\sigma z)^2}\partial_\zeta \dot F + \frac{\tau^3 z^3}{(1-\sigma z)^3}\partial_\zeta^2 F$$
$$+(1-\sigma z)^{1/2} g_{ij}^k (\omega_k F' + z A_k F)(\omega_i\omega_j F'' + z A_{ij} F)\Big\}.$$

Preuve. C'est (3.1.1) pour $\mu = k = 0$, où l'on utilise $t = r(1 - \sigma z)$. □

Nous définirons dorénavant, pour simplifier, l'opérateur H et la forme quadratique q par l'identité

$$\Box u + g\partial u \partial^2 u = \frac{\varepsilon}{r^{5/2}} Q F_0 + \frac{\varepsilon^2}{\sqrt{rt}}\Big\{ HF + q(F,F)\Big\}.$$

Bien entendu, il nous suffira de résoudre l'équation d'onde à des approximations arbitraires $O(z^N) + O(\zeta^{N'})$.

Les résultats de cette étude formelle en z et ζ sont résumées dans la proposition suivante.

PROPOSITION 4.1.1. (i) (*Existence*). *Pour tous* N, N', *et toutes fonctions* $\varphi_\ell(\sigma,\omega)$ $(\ell \le N-2)$, *on peut trouver une fonction* $F = F_0 + \tau G$ *définie pour* $\tau \in [0, \tau_1]$ *telle que* $QF_0 = O(z^N)$, $HF + q(F,F) = O(z^N) + O(\zeta^{N'})$, *avec*

$$\partial_\tau^{2\ell}\partial_z^\ell F(\sigma,\omega,0,0,0) = \varphi_\ell(\sigma,\omega) \quad \textit{pour } 0 \le \ell \le N-2.$$

(ii) (*Unicité*). *Si une fonction* $F = F_0 + \tau G$ *vérifie* $QF_0 = O(z^N)$, $HF + q(F,F) = O(z^N) + O(\zeta^{N'})$ *et* $\partial_\tau^{2\ell}\partial_z^\ell F(\sigma,\omega,0,0,0) = \varphi_\ell(\sigma,\omega)$ *pour* $0 \le \ell \le N-2$, *elle est déterminée à* $O(z^{N-1}) + O(\zeta^{N'})$ *près.*

Preuve. (a) Notons d'abord que $F_0|_{z=0}$ détermine tout le jet en z d'une solution de $QF_0 = 0$, car toutes les fonctions considérées sont bien entendu supportées pour $\sigma \le M$. Inversement, on peut trouver une solution F_0 de $QF_0 = O(z^N)$ pour laquelle $F_0|_{z=0}$ est donnée.

(b) Posons $G = \sum_{k\ge 0} G_k z^k$, $F_0 = \sum_{k\ge 0} f_k z^k$, et ordonnons l'équation $HF + q(F,F) = 0$ selon les puissances de z.

• Pour $z = 0$, on trouve

$$(4.1)_0 \qquad -\tau\dot G_0' - G_0' + g(f_0' + \tau G_0')(f_0'' + \tau G_0'') = 0.$$

• D'autre part, le terme en z^{k-1} dans QG vaut

$$-\Big(\frac14 + \partial_\omega^2\Big) G_{k-1} + 2kG_k' - 2(k-1)G_{k-1}' - (k-1)(k-2)G_{k-1}.$$

Remarquons que dans les opérateurs zA_iF, $zA_{ij}F$, les termes qui contiennent des dérivées en z sont de la forme $z^k\partial_z$, $k \ge 2$, ou $z^4\partial_z^2$; le terme en z^ℓ dans zA_jF

ou $zA_{ij}F$ ne fait donc intervenir que les coefficients f_j, G_j pour $j \leq \ell - 1$. Le terme en z^k dans $q(F,F)$, que nous noterons $q_k(F,F)$, vaut alors

$$
\begin{aligned}
&g\Big\{(f_0' + \tau G_0')(f_k'' + \tau G_k'') + \ldots + (f_k' + \tau G_k')(f_0'' + \tau G_0'')\Big\} \\
&\quad -\frac{\sigma}{2} g\Big\{(f_0' + \tau G_0')(f_{k-1}'' + \tau G_{k-1}'') + \ldots + (f_{k-1}' + \tau G_{k-1}')(f_0'' + \tau G_0'')\Big\} \\
&\quad + g_{ij}^p \omega_i \omega_j \Big\{(f_0'' + \tau G_0'')\,\tilde{A}_p(f_{k-1} + \tau G_{k-1}) + \ldots + (f_{k-1}'' + \tau G_{k-1}'')\,\tilde{A}_p(f_0 + \tau G_0)\Big\} \\
&\quad + g_{ij}^p \omega_p \Big\{(f_0' + \tau G_0')\,\tilde{A}_{ij}(f_{k-1} + \tau G_{k-1}) + \ldots + (f_{k-1}' + \tau G_{k-1}')\,\tilde{A}_{ij}(f_0 + \tau G_0)\Big\} \\
&\quad + \text{termes en } G_j \ \ (j \leq k-2)\ ,
\end{aligned}
$$

où $\tilde{A}_p$ et $\tilde{A}_{ij}$ désignent les parties "de poids zéro" de A_p et A_{ij} (qui ne contiennent pas de dérivées en ζ).

L'annulation du terme en z^k conduit à l'équation

$$
\begin{aligned}
(4.1)_k \quad &-\tau \dot{G}_k' + (2k-1) G_k' + \Big\{ -\Big(\frac{1}{4} + (k-1)(k-2) + \partial_\omega^2\Big) G_{k-1} - 2(k-1) G_{k-1}' \\
&\quad -\frac{1}{4}(\tau \dot{G}_{k-1} + G_{k-1}) + \frac{\tau}{4}(\tau \ddot{G}_{k-1} + 2\dot{G}_{k-1}) - 2\tau^2 \partial_\zeta G_{k-1}' \Big\} \\
&\quad + q_k(F,F) = \text{termes en } F_0\ ,\quad G_j\ \ (j \leq k-2)\ .
\end{aligned}
$$

(c) L'équation $(4.1)_0$ détermine G_0 en fonction de f_0 : en particulier, $G_0 = \frac{g}{2}(f_0')^2 + \frac{g^2}{2}\tau (f_0')^2 f_0'' + \tau^2 \ldots$.

Ecrivons $(4.1)_1$:

$$
-\tau \dot{G}_1' + G_1' - \Big(\frac{1}{2} + \partial_\omega^2\Big) G_0 - 2\tau^2 \partial_\zeta G_0' - 2\tau \partial_\zeta f_0' + q_1(F,F) = 0.
$$

Avec $f_1' = +\frac{1}{2}\Big(\frac{1}{4} + \partial_\omega^2\Big) f_0$, posons $G_1 = \sum_{\ell \geq 0} G_1^\ell \tau^\ell$; pour $\tau = 0$, on obtient de $(4.1)_1$

$$
G_1^{0\prime} - \Big(\frac{1}{2} + \partial_\omega^2\Big) G_0 + \text{termes en } f_0 = 0;
$$

l'annulation du terme en τ dans $(4.1)_1$ fournit

$$
-2\partial_\zeta f_0' + g\Big\{G_0' f_1'' + f_0' G_1^{0\prime\prime} + f_1' G_0'' + f_0'' G_1^{0\prime} - \frac{\sigma}{2}(f_0' G_0'' + f_0'' G_0')\Big\} + \text{termes en } f_0 = 0.
$$

Si l'on exprime G_1^0 dans la deuxième équation à l'aide de la première, on obtient une équation en f_0 de la forme $\partial_\zeta f_0' +$ (termes en f_0 sans dérivées en ζ) $= 0$. Comme $f_0|_{\zeta=0} = \varphi_0$, on en déduit f_0 à $O(\zeta^{N'})$ près.

Avec ce choix de f_0, on résoud $(4.1)_0$ sur un certain intervalle $[0, \tau_1]$, et tous les autres f_k sont déterminés (à la même approximation), ainsi que G_1^0 ; si l'on note $G_1 = G_1^0 + \tau^2 \tilde{G}_1$, l'équation $(4.1)_1$ se réduit, après division par τ^2, à

$$
-\tau \dot{\tilde{G}}_1' - \tilde{G}_1' + g(f_0' + \tau G_0')\tau \tilde{G}_1'' + g(f_0'' + \tau G_0'')\tau \tilde{G}_1' = \text{termes connus}.
$$

Cette équation linéaire possède une (unique) solution sur le même intervalle $[0, \tau_1]$.

On a ainsi résolu $(4.1)_1$, G_1 n'étant déterminé qu'à un élément près du noyau.

(d) Abordons maintenant la résolution de $(4.1)_k$, $k \geq 2$. A ce stade, F_0 est connue, ainsi que les G_j, $j \leq k-2$. Le noyau de l'équation $(4.1)_{k-1}$ est formé des fonctions v vérifiant

$$-\tau \dot{v} + (2k-3)v + \tau g(f_0' + \tau G_0')v' = 0.$$

Ces fonctions sont de la forme

$$v = h\tau^{2k-3} + \tau^{2k-2}\tilde{h},$$

où h est arbitraire, $\tilde{h}$ étant déterminé par h. Plus précisément,

$$v = h\tau^{2k-3} + g f_0' h' \tau^{2k-2} + \frac{1}{2}\Big(g^2(f_0')^2 h'' + (g^2 f_0' f_0'' + g G_0')h'\Big)\tau^{2k-1} + \tau^{2k}\ldots.$$

L'équation $(4.1)_k$ détermine les coefficients G_k^ℓ de $G_k = \sum_{\ell \geq 0} G_k^\ell \tau^\ell$ pour $\ell \leq 2k-4$; les coefficients G_k^{2k-3} et G_k^{2k-2} sont alors connus en fonction de h. L'annulation du terme en τ^{2k-1} conduit à une équation en h de la forme

$$-2\partial_\zeta h' + (\text{termes en } h \text{ sans dérivées en } \zeta) = \text{termes connus}.$$

Comme pour f_0, la connaissance de $h|_{\zeta=0} = \varphi_{k-1}$ détermine h à $O(\zeta^{N'})$ près. On peut alors trouver sur $[0, \tau_1]$ une solution de $(4.1)_k$ sous la forme $G_k = \sum_{0 \leq \ell \leq 2k-2} G_k^\ell \tau^\ell + \tau^{2k}\tilde{G}_k$, $\tilde{G}_k$ satisfaisant une équation linéaire à point singulier régulier et à indices négatifs. □

4.1.2. Choix de u_a^e. Compte tenu de la structure de la solution $U_k = \sum_{\ell \leq k} \varepsilon^\ell u_\ell$ construite à la proposition 3.1, on a

$$\Big(\frac{r^{1/2}}{\varepsilon} U_k\Big)(\sigma, \omega, 0, 0, 0) = \sum_{\ell \leq k} \varepsilon^\ell L_\ell^0(\sigma, \omega, 0) = \varphi_0(\sigma, \omega), \tag{4.1.1}$$

$$(\partial_\tau^{2\ell}\partial_z^\ell)\Big(\frac{r^{1/2}}{\varepsilon} U_k\Big)(\sigma, \omega, 0, 0, 0) = 0.$$

Nous choisissons donc pour u_a^e la fonction $\frac{\varepsilon}{r^{1/2}} F$, où F est donnée par la proposition 4.1, correspondant aux fonctions $\varphi_0, \varphi_1 = 0, \ldots, \varphi_\ell = 0$, φ_0 définie ci-dessus. Comme $(4.1)_0$ signifie que $v = f_0 + \tau G_0$ est solution de $\partial_\tau v = \frac{g}{2}(\partial_\sigma v)^2$, avec $v(\sigma, \omega, 0, \zeta) = f_0(\sigma, \omega, \zeta) = R(\sigma, \omega) + O(\varepsilon) + O(\zeta)$, le temps de vie de G_0 vaut $A_0 + O(\varepsilon)$.

La solution construite $u_a^{II,e}$ existe donc et est régulière pour $0 \leq \tau \leq A$, pour ε assez petit.

4.1.3. Recollement. Dans la période de transition $\varepsilon^{-\lambda} \leq t \leq 2\varepsilon^{-\lambda}$, la solution U_k fournit une autre solution approchée (à $\varepsilon^{k+1-\lambda(\frac{k-3}{2})}$ près) de $\Box u + g\partial u \partial^2 u = 0$ de la forme $\frac{\varepsilon}{r^{1/2}} F$: la partie "unicité" de la proposition 4.1 garantit que les deux solutions diffèrent arbitrairement peu pour N, N' et k assez grands (pour $\lambda < 2$ fixé).

4.2. La construction a l'intérieur. Comme en §4.1, nous nous limitons à $0 \leq \tau \leq A$. Ici, on simplifie résolument la construction à l'extérieur, mais, comme en §3.2, on doit apporter un soin particulier à l'étude du comportement en σ, et au recollement entre les deux périodes.

4.2.1 Les symboles $R(\sigma,\omega)$ et $L(\sigma,\omega)$ ont été définis au 3.2. Définissons $S = S_\varepsilon(\sigma,\omega,\tau)$ comme la solution de

$$\partial_\tau S - \frac{g}{2}(\partial_\sigma S)^2 = 0, \qquad S(\sigma,\omega,0) = R(\sigma,\omega) + \varepsilon L(\sigma,\omega). \tag{4.2.1}$$

Par analogie avec le traitement de [1], nous posons

$$u = u_a^i = \varepsilon u_1^i + \varepsilon^2 z + \chi \frac{\varepsilon}{r^{1/2}} K(r-t,\omega,\tau), \tag{4.2.2}$$

où

$$K(\sigma,\omega,\tau) = S(\sigma,\omega,\tau) - S(\sigma,\omega,0),$$

u_1^i et z ayant été définis au paragraphe 3.2 a et b.

L'approximation obtenue par le choix (4.2.2) est précisée au lemme suivant.

LEMME 4.2.1. *On a* $\left|\partial^\alpha_{x,t,\omega} J_a\right|_0 \le C_\alpha \dfrac{\varepsilon}{t^2}$.

Preuve. (a) Calculons $\Box u$: d'après le lemme 3.2.1 et sa preuve, $\Box z = W$ satisfait à $|\partial^\alpha W|_0 \le C_\alpha/(1+t)^{3/2}$. Donc

$$\Box u = \varepsilon^2 W - \frac{\varepsilon^2 \chi}{\sqrt{rt}} \partial_\tau K' + H_1,$$

où

$$H_1 = \frac{\varepsilon}{r^{1/2}}\left\{(\chi_{tt} - \chi'')K - 2K'(\chi_t + \chi') + \frac{\varepsilon}{\sqrt{t}}\partial_\tau K\left(\chi_t - \frac{\chi}{4t}\right) + \chi\frac{\varepsilon^2}{4t}\partial_\tau^2 K - \frac{\chi}{r^2}\left(\frac{1}{4} + \partial_\omega^2\right)K\right\}.$$

Par ailleurs, on peut écrire u sous la forme

$$\begin{aligned} u &= \chi\frac{\varepsilon S(r-t,\omega,\tau)}{r^{1/2}} + \varepsilon\left(u_1 - \chi\frac{R(r-t,\omega)}{r^{1/2}}\right) + \varepsilon^2\left(z - \chi\frac{L(r-t,\omega)}{r^{1/2}}\right) \\ &= \chi\frac{\varepsilon}{r^{1/2}} S(r-t,\omega,\tau) + \varepsilon r_1, \end{aligned}$$

les estimations de r_1 résultant du lemme 3.2.b. On a donc

$$g_{ij}^k \partial_k u \partial_{ij}^2 u = \varepsilon^2 g \frac{\chi^2}{r} S' S'' + \varepsilon^2 H_2,$$

avec

$$\begin{aligned} H_2 = &\left\{g_{ij}^k \partial_k\left(\frac{\chi S}{r^{1/2}}\right)\partial_{ij}^2\left(\chi\frac{S}{r^{1/2}}\right) - g\frac{\chi^2}{r} S' S''\right\} \\ &+\left\{g_{ij}^k \partial_k\left(\chi\frac{S}{r^{1/2}}\right)\partial_{ij}^2 r_1 + g_{ij}^k \partial_k r_1 \partial_{ij}^2\left(\chi\frac{S}{r^{1/2}}\right) + g_{ij}^k \partial_k r_1 \partial_{ij}^2 r_1\right\} = H_2^1 + H_2^2. \end{aligned}$$

Finalement

$$J_a = \varepsilon^2 W + H_1 + \varepsilon^2 H_2 + \chi\frac{\varepsilon^2}{r\sqrt{t}}\partial_\tau S' \frac{t-r}{\sqrt{t+\sqrt{r}}} + \frac{\chi(\chi-1)}{r}\varepsilon^2 g S' S''.$$

(b) On voit facilement, comme dans [1], que S est borné en σ, tandis que S' est essentiellement en $\frac{1}{\sigma}$ ($|\sigma| \to +\infty$). On en déduit que $K = \frac{g}{2}\int_0^\tau R'^2(\sigma,\omega,s)\,ds$ se comporte en $\frac{1}{\sigma^2}$. D'autre part, à l'intérieur, la fonction S ne présente aucune singularité pour $\tau \leq A$, à cause de (ND). On a alors

$$|\partial^\alpha H_1|_0 \leq \frac{C\varepsilon}{(1+t)^2} \quad \text{et} \quad |\partial^\alpha H_2^1|_0 \leq \frac{C}{(1+t)^{3/2}}.$$

Un terme typique de H_2^2 est

$$\partial_{ij}^2\Big(\chi\frac{S}{r^{1/2}}\Big)\Big\{\partial_k\Big(u_1 - \chi\frac{R}{r^{1/2}}\Big) + \varepsilon\partial_k\Big(z - \chi\frac{L}{r^{1/2}}\Big)\Big\} :$$

le premier produit est majoré par $\dfrac{S_{-3}}{t^2}$, le second par $\varepsilon\dfrac{S_{-5/2}}{t^{3/2}}$; ce terme est donc majoré en norme L^2 par

$$\frac{C}{(1+t)^{3/2}} + \frac{C\varepsilon}{1+t} \leq \frac{C}{(1+t)^{3/2}};$$

quant au terme $\partial_k r_1\,\partial_{ij}^2 r_1$, il est majoré par

$$\frac{S_{-2}}{t^3} + \varepsilon\frac{S_{-1}}{t^{5/2}} + \varepsilon^2\frac{S_{-1}}{t^2}.$$

Au total, $|\partial^\alpha H_2^2|_0 \leq C/(1+t)^{3/2}$. Les autres termes de J_a sont négligeables, et l'on trouve

$$|\partial^\alpha J_a|_0 \leq \frac{C\varepsilon^2}{(1+t)^{3/2}} + \frac{C\varepsilon}{(1+t)^2} \leq C\frac{\varepsilon}{(1+t)^2}. \qquad \square$$

4.2.2. Le recollement. Les solutions u_a^i construites en périodes I et II aux §§3.2 et 4.2.1. se recollent très bien en période de transition $\varepsilon^{-\lambda} \leq t \leq 2\varepsilon^{-\lambda}$. En effet, on peut écrire

$$\begin{aligned} u^I = \varepsilon u_1 + \chi\frac{g}{2r^{1/2}}\varepsilon\tau R'^2 + \varepsilon^2 z + \frac{\chi g^2}{6r^{1/2}}\varepsilon\tau^2(R'^3)' + \frac{\chi g}{r^{1/2}}\varepsilon^2\tau R'L' \\ + \chi\frac{g^3}{6r^{1/2}}\varepsilon\tau^3(3R'^2R''^2 + R'^3R''') + \chi\frac{g^2}{2r^{1/2}}\varepsilon^2\tau^2(2R'R''L' + R'^2L'') \\ + \varepsilon^5\overline{u}_5 + \varepsilon^6\overline{u}_6 + \varepsilon^7\overline{u}_7, \end{aligned}$$

et

$$\begin{aligned} u^{II} - u^I = \chi\frac{\varepsilon}{r^{1/2}}\Big\{S(r-t,\omega,\tau) - R - \frac{g}{2}\tau R'^2 - \frac{g^2}{6}\tau^2(R'^3)' \\ - \frac{g^3}{6}\tau^3(3R'^2R''^2 + R'^3R''') - a_5^1\tau^4 - a_6^1\tau^5 - a_7^1\tau^6\Big\} \\ - \chi\frac{\varepsilon^2}{r^{1/2}}\Big\{L + g\tau R'L' + \frac{g^2}{2}\tau^2(2R'R''L' + R'^2L'')\Big\}. \end{aligned}$$

Il faut d'abord éclaircir la structure des termes principaux $\overline{u}_\ell = \chi a_\ell^1(r-t,\omega t^{\frac{1}{2}(\ell-2)})$ de u_ℓ, en affinant l'analyse amorcée au §3.2.4. ($a_1^1(\sigma,\omega) = R(\sigma,\omega)$).

LEMME 4.2.2. *L'expression* $\sum_{\ell\geq 1}\tau^{\ell-1}a^1_\ell(\sigma,\omega)$ *est le développement de Taylor en* τ *de la solution* $R(\sigma,\omega,\tau)$ *de l'équation*

$$\partial_\tau R=\frac{g}{2}(\partial_\sigma R)^2,\qquad R(\sigma,\omega,0)=R(\sigma,\omega). \tag{4.2.3}$$

Preuve. (a) Le terme principal de $Q_\ell=\sum_{\ell'+\ell''=\ell}g^k_{ij}\,\partial_k u_{\ell'}\,\partial^2_{ij}u_{\ell''}$ vaut

$$\chi t^{\frac{1}{2}(\ell-4)}g\sum(a^1_{\ell'})'(a^1_{\ell''})''=\chi g t^{\frac{1}{2}(\ell-4)}\Big\{\sum_{\ell'<\ell'',\,\ell'+\ell''=\ell}(a^1_{\ell'})'(a^1_{\ell''})'+\frac{1}{2}[(a^1_{\ell/2})']^2\Big\}',$$

d'où, pour $\ell\geq 2$ d'après (3.2.1),

$$a^1_\ell=\frac{1g}{\ell-1}\Big\{\quad\Big\}.$$

(b) Pour une donnée arbitraire $A_1(\sigma,\omega)$, soit $A(\sigma,\omega,\tau)$ la solution de $\partial_\tau A=\frac{g}{2}(A')^2$, $A(\sigma,\omega,0)=A_1(\sigma,\omega)$. Elle s'écrit formellement

$$A=A_1+\tau A_2+\tau^2A_3+\cdots+\tau^kA_{k+1}+\cdots,$$

en sorte que l'équation

$$A_2+2\tau A_3+\cdots+k\tau^{k-1}A_{k+1}+\cdots=\frac{g}{2}\Big\{A'_1+\tau A'_2+\cdots+\tau^kA'_{k+1}+\cdots\Big\}^2$$

conduit aux relations

$$A_2=\frac{g}{2}(A'_1)^2,\ldots,(k-1)A_k=\frac{g}{2}\Big\{2\sum_{\substack{\ell'+\ell''=k\\ \ell'\leq\ell''}}A'_{\ell'}A'_{\ell''}+(A'_{k/2})^2\Big\}.$$

Comme $a^1_1=R(\sigma,\omega)$, le lemme est prouvé. □

De la même façon, en notant $\bar{\bar{u}}_\ell=\chi a^2_\ell(r-t,\omega)t^{\frac{1}{2}(\ell-3)}$ le terme "sous principal" de u_ℓ ($a^2_1=0$, $a^2_2=L$), on obtient le lemme suivant.

LEMME 4.2.3. *L'expression* $\sum_{\ell\geq 2}\tau^{\ell-2}a^2_\ell(\sigma,\omega)$ *est le développement de Taylor en* τ *de la solution* $L(\sigma,\omega,\tau)$ *de l'équation*

$$\partial_\tau L=g(\partial_\sigma R)(\partial_\sigma L),\qquad L(\sigma,\omega,0)=L(\sigma,\omega), \tag{4.2.4}$$

la fonction $R(\sigma,\omega,\tau)$ *étant définie par* (4.2.3).

Preuve. (a) Le terme "sous principal" de Q_ℓ est le terme principal dans l'expression

$$\sum_{\ell'+\ell''=\ell}\Big\{g^k_{ij}\,\partial_k\bar{u}_\ell,\,\partial^2_{ij}\bar{\bar{u}}_{\ell''}+g^k_{ij}\,\partial_k\bar{\bar{u}}_{\ell'}\,\partial^2_{ij}\bar{u}_{\ell''}\Big\},$$

soit

$$\chi t^{\frac{1}{2}(\ell-5)}g\sum_{\ell'+\ell''=\ell}\Big\{(a^1_{\ell'})'(a^2_{\ell''})''+(a^2_{\ell'})'(a^1_{\ell''})''\Big\}.$$

Donc, pour $\ell\geq 2$, grâce à (3.2.1),

$$a^2_\ell=\frac{g}{\ell-2}\Big\{\sum_{\substack{\ell'<\ell''\\ \ell'+\ell''=\ell}}(a^1_{\ell'})'(a^2_{\ell''})'+(a^1_{\ell''})'(a^2_{\ell'})'+(a^1_{\ell/2})'(a^2_{\ell/2})'\Big\}.$$

(b) Pour une donnée arbitraire $B_2(\sigma,\omega)$, soit $B(\sigma,\omega,\tau)$ la solution de

$$\partial_\tau B = g A' B', \qquad B(\sigma,\omega,0) = B_2,$$

où $A = A_1 + \tau A_2 + \cdots$.

Elle s'écrit $B = B_2 + \tau B_3 + \cdots + \tau^k B_{k+2} + \cdots$, en sorte que l'équation

$$B_3 + \cdots + k\tau^{k-1} B_{k+2} + \cdots = g(A_1' + \cdots + \tau^k A_{k+1}' + \cdots)(B_2' + \cdots + \tau^j B_{j+2}' + \cdots)$$

conduit aux relations

$$(\ell - 2) B_\ell = g \sum_{\ell'+\ell''=\ell} A_{\ell'}' B_{\ell''}'. \qquad \square$$

Compte tenu des lemmes 4.2.2 et 4.2.3, on peut donc écrire

$$\begin{aligned} u^{II} - u^I = \chi \frac{\varepsilon}{r^{1/2}} \Big\{ S(r-t,\omega,\tau) - R(r-t,\omega,\tau) - \varepsilon L(r-t,\omega,\tau) \Big\} \\ + \chi \frac{\varepsilon}{r^{1/2}} \Big\{ R(r-t,\omega,\tau) - \sum_{1\le\ell\le 7} \tau^{\ell-1} a_\ell^1(r-t,\omega) \Big\} \\ + \chi \frac{\varepsilon}{r^{1/2}} \Big\{ L(r-t,\omega,\tau) - \sum_{2\le\ell\le 4} \tau^{\ell-2} a_\ell^2(r-t,\omega) \Big\} . \end{aligned}$$

Le premier terme est étudié au lemme suivant.

LEMME 4.2.4. *La solution $S(\sigma,\omega,\tau)$ de (4.2.1) s'écrit*

$$S(\sigma,\omega,\tau) = R(\sigma,\omega,\tau) + \varepsilon L(\sigma,\omega,\tau) + \varepsilon^2 \tilde{S}(\sigma,\omega,\tau),$$

où R et L sont définis respectivement par (4.2.3) et (4.2.4), et où $\tilde{S}$ est intégrable en σ.

Preuve. En remplaçant dans l'équation de S, on trouve

$$\begin{aligned} \partial_\tau R + \varepsilon \partial_\tau L + \varepsilon^2 \partial_\tau \tilde{S} = \frac{g}{2} (\partial_\sigma R + \varepsilon \partial_\sigma L + \varepsilon^2 \partial_\sigma \tilde{S})^2 = \frac{g}{2} (\partial_\sigma R)^2 + \varepsilon g (\partial_\sigma R)(\partial_\sigma L) \\ + \varepsilon^2 g \Big(\partial_\sigma R \partial_\sigma \tilde{S} + \frac{1}{2} (\partial_\sigma L)^2 \Big) + \varepsilon^3 g \partial_\sigma L \partial_\sigma \tilde{S} + \frac{g}{2} \varepsilon^4 (\partial_\sigma \tilde{S})^2, \end{aligned}$$

soit

$$\partial_\tau \tilde{S} = g(\partial_\sigma R + \varepsilon \partial_\sigma L) \partial_\sigma \tilde{S} + \varepsilon^2 \frac{g}{2} (\partial_\sigma \tilde{S})^2 + \frac{g}{2} (\partial_\sigma L)^2, \qquad \tilde{S}(\sigma,\omega,0) = 0.$$

On voit facilement (comme dans [1, Lem. 5.1.1]) que $R(\sigma,\omega,\tau)$ se comporte comme un symbole d'ordre $-\frac{1}{2}$ en σ, tandis que $\partial_\sigma L$ est dans $L^2(\mathbb{R}_\sigma)$. On en déduit que $\tilde{S}$ est intégrable en σ. $\square$

On obtient finalement les estimations suivantes:

$$\begin{aligned} \big| \partial_{x,t,\omega}^\alpha (u^{II} - u^I) \big|_0 &\le C\varepsilon^3 + C\varepsilon^8 (1+t)^{7/2} + C\varepsilon^5 (1+t)^{3/2} \\ &\le C\varepsilon^8 (1+t)^{7/2} \end{aligned} \tag{4.2.5}$$

(car $t \sim \varepsilon^{-\lambda}$, $\lambda > \frac{3}{2}$),

$$\big\| \partial_{x,t,\omega}^\alpha (u^{II} - u^I) \big\|_0 \le C\varepsilon^8 (1+t)^3. \tag{4.2.6}$$

4.2.3. Estimation de J_a en zone de transition et choix de λ. On pose, pour $\varepsilon^{-\lambda} \le t \le 2\varepsilon^{-\lambda}$,

$$u = u_a = \theta(t\varepsilon^{+\lambda})u^I + (1-\theta(t\varepsilon^{+\lambda}))u^{II}.$$

LEMME 4.2.5. *En zone de transition, on a, si* $\frac{3}{2} < \lambda \le \frac{14}{9}$,

$$|\partial^\alpha_{x,t,\omega} J_a|_0 \le \frac{C_\alpha \varepsilon}{t^2}.$$

Preuve. On écrit

$$\begin{aligned} J_a = \theta\Box u^I + (1-\theta)\Box u^{II} + 2\varepsilon^{+\lambda}\theta'(u^I - u^{II}) + \varepsilon^{+2\lambda}\theta''(u^I - u^{II}) \\ + g^k_{ij}\theta^2 \partial_k u^I \partial^2_{ij} u^I + g^k_{ij}(1-\theta)^2 \partial_k u^{II} \partial^2_{ij} u^{II} \\ + \theta(1-\theta) g^k_{ij}(\partial_k u^I \partial^2_{ij} u^{II} + \partial_k u^{II} \partial^2_{ij} u^I) + r \ , \end{aligned}$$

où r regroupe les termes contenant au moins une dérivée de θ, d'où

$$|\partial^\alpha r|_0 \le C \frac{\varepsilon^{2+\lambda}}{(1+t)^{1/2}}.$$

Donc

$$\begin{aligned} J_a = \theta J^I_a + (1-\theta) J^{II}_a + r + (2\varepsilon^{+\lambda}\theta' + \varepsilon^{+2\lambda}\theta'')(u^I - u^{II}) \\ + \theta(\theta-1)\Big\{ g^k_{ij}\partial_k u^I \partial^2_{ij} u^I + g^k_{ij}\partial_k u^{II}\partial^2_{ij} u^{II} \\ - g^k_{ij}(\partial_k u^I \partial^2_{ij} u^{II} + \partial_k u^{II}\partial^2_{ij} u^I)\Big\} \end{aligned}$$

et le dernier terme vaut $\theta(\theta-1) g^k_{ij}\partial_k(u^I - u^{II})\partial^2_{ij}(u^I - u^{II})$. Finalement,

$$\begin{aligned} |\partial^\alpha J_a|_0 \le C\Big\{ \frac{\varepsilon^3}{1+t} + \varepsilon^5(1+t)^{1/2} + \varepsilon^8(1+t)^{5/2} + \frac{\varepsilon}{(1+t)^2} + \frac{\varepsilon^{2+\lambda}}{(1+t)^{1/2}} \\ + \varepsilon^{+\lambda+8}(1+t)^{7/2} + \varepsilon^{16}(1+t)^{6,5}\Big\} \le C\frac{\varepsilon}{(1+t)^2} \end{aligned}$$

pour $\frac{3}{2} \le \lambda \le \frac{14}{9}$. □

On fixe dorénavant $\lambda = \frac{14}{9}$.

4.3. Le recollement intérieur/extérieur. On procède exactement comme au §3.3.

• On a

$$u^e_a = \frac{\varepsilon}{r^{1/2}} F\Big(r-t, \omega, \frac{1}{r}, \tau, \zeta\Big),$$

et avec les notations de la proposition 4.1.1, $F = f_0 + \tau G_0 + O(z)$. D'après $(4.1)_0$, la fonction $v = f_0 + \tau G_0$ est solution de

$$\dot v = \frac{g}{2}(\partial_\sigma v)^2, \qquad v(\sigma,\omega,0,\zeta) = f_0(\sigma,\omega,\zeta), \tag{4.3.1}$$

et la fonction f_0 satisfait, par construction et grâce à (4.1.1),

$$f_0 = \varphi_0 + O(\zeta) = R(\sigma,\omega) + \varepsilon L(\sigma,\omega) + \sum_{2\le\ell\le k} \varepsilon^\ell L^0_\ell(\sigma,\omega,0) + O(\zeta). \tag{4.3.2}$$

On a donc

$$\|\partial^\alpha (S-v)\|_0 \leq C_\alpha \varepsilon^2 (\log t)\ . \tag{4.3.3}$$

• Par ailleurs, dans la zone de recollement,

$$u_1 - \frac{R}{r^{1/2}} \quad \text{et} \quad z - \frac{L}{r^{1/2}}$$

se comportent comme les premiers termes négligés qui sont de la forme

$$\frac{S_{1/2}(r-t,\omega)}{r^{3/2}} \quad \text{et} \quad \frac{S_{-2}(r-t,\omega)}{r},$$

respectivement. Donc

$$\left\| (\partial_t + \partial_r)\left[\varepsilon\Big(u_1 - \frac{R}{r^{1/2}}\Big) + \varepsilon^2\Big(z - \frac{L}{r^{1/2}}\Big)\right]\right\|_0 \leq C\frac{\varepsilon}{t^{5/2}}.$$

La preuve est alors identique à celle de la proposition 3.3.1 et conduit au résultat suivant.

PROPOSITION 4.3.1. *Pour $u_a = u_a^{II} = \theta u_a^i + (1-\theta)u_a^e$ on a*

$$\left|\partial^\alpha_{x,t,\omega} J_a\right|_0 \leq C_\alpha \frac{\varepsilon \log t}{t^2}.$$

5. Conclusions pour les périodes I et II. On a fixé A, $0 < A < A_0$ et l'on a construit aux paragraphes 3 et 4 une solution approchée u_a (pour $t \leq \dfrac{A^2}{\varepsilon^2}$) de l'équation $\Box u + g\,\partial u\,\partial^2 u = 0$ sous la forme

$$u_a = \theta(t\varepsilon^\lambda)\Big\{\theta(r-t+C_0-1)u_a^{I,i} + (1-\theta(r-t+C_0-1))u_a^{I,e}\Big\}$$
$$+(1-\theta)(t\varepsilon^\lambda)\Big\{\theta(r-t+C_0-1)u_a^{II,i} + (1-\theta(r-t+C_0-1))u_a^{II,e}\Big\}\Big(\lambda = \frac{14}{9}\Big),$$

pour laquelle on a prouvé l'estimation

$$\int_0^{\frac{A^2}{\varepsilon^2}} \left|\partial^\alpha_{x,t,\omega} J_a\right|_0 ds \leq C_\alpha \varepsilon^{23/9}\,|\log\varepsilon| \tag{5.1}$$

(c'est une conséquence immédiate des propositions 3.3, 4.3 et du lemme 4.2.5).

En fait, pour tout $p \in \mathbb{N}$, on peut choisir q en sorte que u_a vérifie, à l'extérieur,

$$\left\|\partial^\alpha_{x,t,\omega} J_a\right\|_0 \leq C_\alpha \varepsilon^p. \tag{5.2}$$

Un argument standard d'inégalités d'énergie (cf. [5], [1]), utilisant l'hypothèse d'induction sur le temps

$$\left\|\partial^2_{x,t,\omega}(u-u_a)\right\|_0 \leq \frac{\varepsilon}{(1+t)^{1/2}},$$

permet alors de conclure que, pour $0 < A < A_0$ et ε petit, on a, pour $t \leq \dfrac{A^2}{\varepsilon^2}$,

$$\left|\partial^\alpha_{x,t,\omega} \nabla(u-u_a)\right|_0 \leq C_\alpha \varepsilon^{23/9}\,|\log\varepsilon|\,. \tag{5.3}$$

Le même argument, utilisé cette fois dans un domaine $t \geq C_0$, $t - C_0 \leq r \leq 2\dfrac{A^2}{\varepsilon^2} + M - t$, où seule (5.2) intervient, montre qu'en fait, pour tout p, on peut trouver q tel que

$$\text{(5.4)} \qquad \text{A l'extérieur,} \quad \left\| \partial^{\alpha}_{x,t,\omega} \nabla (u - u_a) \right\|_0 \leq C_{\alpha}\, \varepsilon^p.$$

6. La solution en période *III*. Nous n'avons pour l'instant étudié u_a et u que pour $\tau \leq A$, $0 < A < A_0$.

Nous abordons maintenant l'étude du moment où τ s'approche de la valeur critique $\tau_*(\varepsilon)$, où les dérivées secondes de u_a deviennent infinies quelque part à l'extérieur (cf. §2).

A l'intérieur, nous conserverons la solution approchée $u_a^{II,i}$ construite au paragraphe 4.2, puisqu'elle n'explose pas.

6.1. Nous allons d'abord simplifier la structure de $u_a^{II,e}$ en remarquant que, pour

$$\tau \geq A > 0, \qquad z = \frac{\varepsilon^2}{\tau^2}\left(1 + \frac{\sigma \varepsilon^2}{\tau^2}\right)^{-1}.$$

Le lemme suivant est une variante du lemme 4.1.

LEMME 6.1.1. *Pour toute fonction F, en posant $u = \dfrac{\varepsilon}{r^{1/2}} F(r - t, \omega, \tau)$, on a*

$$\text{(6.1.1)} \qquad \Box u + g\, \partial u\, \partial^2 u$$
$$= \frac{\varepsilon^2}{\sqrt{rt}} \Bigg\{ - \dot{F}' + g F' F'' + \varepsilon^2 \Bigg[- \frac{1}{\tau^3}\left(\frac{1}{4} + \partial_{\omega}^2\right) F - \frac{\sigma g}{2\tau^2} F' F'' - \frac{\dot{F}}{4\tau^2} + \frac{\ddot{F}}{4\tau}$$
$$+ g^k_{ij}\left(\omega_k F' A_{ij} F' + \omega_i \omega_j F'' A_k F\right)\Bigg] + \varepsilon^4 R(F) \Bigg\},$$

où les A_k, A_{ij} sont des opérateurs différentiels en ∂_{ω}, ∂_{τ} d'ordre 1 à coefficients réguliers en $(\omega, \tau, \varepsilon^2\sigma)$, et R est une expression quadratique de dérivées de F précisée en (6.2.1).

Preuve. (a) D'après (3.1.1), on a

$$\Box u = \frac{-\varepsilon}{r^{5/2}}\left(\frac{1}{4} + \partial_{\omega}^2\right) F - \frac{\varepsilon^2}{\sqrt{rt}}\left(\dot{F}' + \frac{\dot{F}}{4r}\right) + \frac{\varepsilon^3}{\sqrt{r}\, t}\, \frac{\ddot{F}}{4}.$$

(b) On a, pour $i \geq 1$,

$$\tilde{\partial}_i F = r^{1/2}\, \partial_i \frac{F}{r^{1/2}} = \omega_i F' + \frac{\varepsilon^2}{\tau^2}\left(\frac{t}{r}\right)\left(- \frac{\omega_i}{2} F + \omega_i \perp \partial_{\omega} F\right)$$

et

$$\tilde{\partial}_0 F = \partial_0 F = -F' + \frac{\varepsilon^2}{2\tau} \dot{F},$$

donc $\tilde{\partial}_i = \omega_i \partial_{\sigma} + \varepsilon^2 A_i$, où A_i est un opérateur du premier ordre en $\partial_{\omega}, \partial_{\tau}$, à coefficients réguliers en $(\omega, \tau, \varepsilon^2\sigma)$, car $\dfrac{t}{r} = \left(1 + \varepsilon^2 \dfrac{\sigma}{\tau^2}\right)^{-1}$.

De façon générale, pour $|\alpha| = k$,

$$\tilde{\partial}^{\alpha} = \omega^{\alpha}\, \partial_{\sigma}^k + \varepsilon^2 A^1_{\alpha}\, \partial_{\sigma}^{k-1} + \varepsilon^4 A^2_{\alpha}\, \partial_{\sigma}^{k-2} + \cdots + \varepsilon^{2k} A^k_{\alpha},$$

où A^k_α est un opérateur d'ordre k en $(\partial_\omega, \partial_\tau)$, à coefficients réguliers en $(\omega, \tau, \varepsilon^2\sigma)$.

On notera en particulier $\tilde\partial^2_{ij} F = \omega_i\omega_j F'' + \varepsilon^2 A_{ij} F' + \varepsilon^4 B_{ij} F$, en sorte que les termes quadratiques $g\partial u\partial^2 u$ s'écrivent

$$\frac{1}{\varepsilon^2} g\partial u\partial^2 u = g^k_{ij}\frac{1}{r}\tilde\partial_k F\tilde\partial^2_{ij} F = \frac{1}{r}\Big\{ g F' F'' + \varepsilon^2 g^k_{ij}(A_k F\omega_i\omega_j F'' + \omega_k F' A_{ij} F') + \varepsilon^4 g^k_{ij}(A_k F A_{ij} F' + \omega_k F' B_{ij} F) + \varepsilon^6 g^k_{ij} A_k F B_{ij} F\Big\}.$$

On obtient ainsi le résultat, avec

$$R(F) = \Big(\frac{t}{r}\Big)^{1/2}\Big\{ g^k_{ij}(A_k F A_{ij} F' + \omega_k F' B_{ij} F) + \varepsilon^2 g^k_{ij} A_k F B_{ij} F\Big\} + * g^k_{ij}(\omega_k F' A_{ij} F' + \omega_i\omega_j F'' A_k F) + \sum_{|\alpha|\le 2} * \partial^\alpha_{\sigma,\omega,\tau} F, \tag{6.1.2}$$

$*$ désignant des coefficients réguliers en (σ, ω, τ). □

Dorénavant, nous noterons $E(F)$ le terme $-\frac{1}{\tau^3}\big(\frac{1}{4} + \partial^2_\omega\big) F + \cdots$ entre crochets au lemme 6.1.1.

Désignons par $F_A(\sigma, \omega)$ la valeur prise en $\tau = A$ par $\dfrac{r^{1/2}}{\varepsilon} u^e_a$, construite au §4.1.

Notons alors F la solution de

$$\partial_\tau F = \frac{g}{2}(\partial_\sigma F)^2, \qquad F(\sigma, \omega, A) = F_A(\sigma, \omega), \tag{6.1.3}$$

et G la solution de

$$\partial_\tau G - g\partial_\sigma F\partial_\sigma G = \int_M^\sigma E(F)\, ds, \qquad G(\sigma, \omega, A) = 0. \tag{6.1.4}$$

Enfin, posons

$$u_a = u^{III,e}_a = \frac{\varepsilon}{r^{1/2}}\Big(F(r-t, \omega, \tau) + \varepsilon^2 G(r-t, \omega, \tau)\Big).$$

6.2. Estimation de l'erreur sur $\tau = A$. Le raccord (brutal) entre les périodes *II* et *III* est justifié par le lemme suivant.

LEMME 6.2.1. *La fonction S étant définie par* (4.2.1), *on a*

(i) $F_A(\sigma, \omega) = S(\sigma, \omega, A) + O(\varepsilon^2 \log \varepsilon)$,

(ii) *Pour $\tau = A$, $\big\|\partial^\alpha_{x,\omega}(\partial^+_t)^k(u - u_a)\big\|_0 \le C_{\alpha,k}\varepsilon^8$, où ∂^+_t désigne la dérivée à droite en $t = \dfrac{A^2}{\varepsilon^2}$.*

Preuve. (a) L'estimation (i) est prouvée au §4.3.

(b) Notons $u^I_a = \dfrac{\varepsilon}{r^{1/2}} F_\pm(r - t, \omega, \tau)$ les solutions approchées en périodes *II* et *III* (pour $t \le A^2/\varepsilon^2$ et $t \ge A^2/\varepsilon^2$, respectivement), exprimées en variables (σ, ω, τ) au voisinage de $\tau = A$.

On a, pour

$$\tau = A, \qquad \partial^+_t(u - u_a) = \partial_t u - \frac{\varepsilon}{r^{1/2}}\Big(-F'_+ + \frac{\varepsilon}{2\sqrt{t}}\dot F_+\Big),$$

$$\partial_t^-(u-u_a) = \partial_t u - \frac{\varepsilon}{r^{1/2}}\Big(-F'_- + \frac{\varepsilon}{2\sqrt{t}}\dot{F}_-\Big),$$

en sorte que

$$\partial_t^+(u-u_a) - \partial_t^-(u-u_a) = \frac{\varepsilon}{r^{1/2}}\Big((F_+ - F_-)' - \frac{\varepsilon}{2\sqrt{t}}\overbrace{(F_+ - F_-)}^{\bullet}\Big) = -\frac{\varepsilon^2}{2\sqrt{rt}}\overbrace{(F_+ - F_-)}^{\bullet},$$

à cause des choix (6.1.3), (6.1.4) ($F_+ = F + \varepsilon^2 G$).

Or, si une fonction régulière F satisfait une équation du type

$$-\dot{F}' + gF'F'' + \varepsilon^2[\text{dérivées de } F] + \cdots = r,$$

la donnée $F|_{\tau=A}$ détermine $\dot{F}|_{\tau=A}$ à une erreur près de l'ordre de r.

Pour F_-, r est arbitrairement petit, tandis que $r = O(\varepsilon^4)$ pour F_+ ; donc

$$\overbrace{(F_+ - F_-)}^{\bullet} = O(\varepsilon^4).$$

Comme d'autre part on sait par (5.4) que $\big\|\partial_{x,\omega}^{\alpha}(\partial_t^-)^k(u-u_a)\big\|_0 \le C_\alpha \varepsilon^p$, on obtient (ii). □

6.3. Estimations des dérivées de F, G et u_a. La fonction $v = -gF'$ est solution de l'équation de Burger.

6.3.1. Solutions de l'équation de Burger.

LEMME 6.3.1. *Soit $v = v(\sigma,\omega,\tau)$ la solution de l'équation de Burger*

$$\partial_\tau v + v\partial_\sigma v = 0, \qquad v(\sigma,\omega,0) = w(\sigma,\omega), \tag{6.3.1}$$

où $w \in C^\infty$ est supposée telle que $\inf_{\sigma,\omega} \partial_\sigma w = \dfrac{-1}{\tau_*} < 0$. Pour $0 \le \tau < \tau_*$, on peut définir $X = X(\sigma,\omega,\tau)$ et $D = D(X,\omega,\tau)$ par

$$X + \tau w(X,\omega) = \sigma, \qquad D(X,\omega,\tau) = 1 + \tau(\partial_\sigma w)(X,\omega). \tag{6.3.2}$$

Alors toute dérivée $\partial_\sigma^i \partial_\omega^j \partial_\tau^\ell v$ $(i+j+\ell = k)$ est de la forme

$$\partial_\sigma^i \partial_\omega^j \partial_\tau^\ell v(\sigma,\omega,\tau) = \sum_{\substack{0 \le 2p \le k-1+q \\ q \le k-1}} a_{pq}(X,\omega,\tau)\frac{(\partial_\sigma^2 w)^q}{D^{k+p}}(X,\omega,\tau), \tag{6.3.3}$$

pour certains coefficients a_{pq} réguliers.

En particulier,

$$\Big|\partial_\sigma^i \partial_\omega^j \partial_\tau^\ell v(\sigma,\omega,\tau)\Big| \le \frac{C}{D^{3k/2-1/2}}\ . \tag{6.3.4}$$

Preuve. (a) Il est bien connu que $v(\sigma,\omega,\tau) = w(X,\omega)$, d'où $\partial_\sigma v = (\partial_\sigma w)\partial_\sigma X$, $\partial_\omega v = (\partial_\sigma w)\partial_\omega X + \partial_\omega w$, $\partial_\tau v = (\partial_\sigma w)\partial_\tau X$, avec

$$\partial_\sigma X = \frac{1}{D}, \quad \partial_\omega X = -\tau\frac{\partial_\omega w}{D}, \quad \partial_\tau X = -\frac{w}{D},$$

ce qui prouve (6.3.3) pour $k = 1$.

(b) La dérivation (en σ par exemple) d'un terme $a_{pq}(\partial_\sigma^2 w)^q / D^{k+p}$ produit une somme de la forme

$$\partial_X a_{pq} \frac{(\ )^q}{D^{k+p+1}} + \frac{a_p}{D^{k+p}} q(\ )^{q-1} \frac{(\partial_\sigma^3 w)}{D} - (k+p) a_{pq} \frac{(\ )^q}{D^{k+p+1}} \tau(\partial_\sigma^2 w) \frac{1}{D},$$

d'où (6.3.3) par récurrence.

(c) Comme $D \geq 0$, on a $|\nabla_{X,\omega} D| \leq \text{cte } D^{1/2}$, soit $\tau(|\partial_\sigma^2 w| + |\partial_\omega \partial_\sigma w|) \leq \text{cte } D^{1/2}$. Comme $D \geq \text{cte} > 0$ pour $\tau \geq 0$ petit, on en déduit $|\partial_\sigma^2 w| + |\partial_\omega \partial_\sigma w| \leq \text{cte } D^{1/2}$.

La majoration (6.3.4) découle alors immédiatement de (6.3.3). □

6.3.2. L'équation de Burger inhomogène.

LEMME 6.3.2. *Soit $h(\sigma, \omega, \tau)$ la solution de l'équation*

$$\partial_\tau h + v \partial_\sigma h + h \partial_\sigma v = \psi(X, \omega, \tau), \qquad h(\sigma, \omega, 0) = 0 \ , \tag{6.3.5}$$

où v et X sont définis au lemme 6.3.1, et ψ est donnée.

Alors, pour $\tau < \tau_*$, toute dérivée $\partial_\sigma^i \partial_\omega^j \partial_\tau^\ell h$ $(i + j + \ell = k)$ est de la forme

$$\partial_\sigma^i \partial_\omega^j \partial_\tau^\ell h = \sum_{\substack{\ell'' \leq k \\ \ell' \leq 2(k-\ell'') \\ q \geq 2\ell' - 3(k-\ell'')}} *(X, \omega, \tau) \frac{(\partial_\sigma^2 w)^q}{D^{1+\ell'+\ell''}} \partial_{X,\omega,\tau}^{\ell''} C, \tag{6.3.6}$$

où $C(X, \omega, \tau) = \int_0^\tau (D\psi)(X, \omega, s)\, ds$.

En particulier,

$$\left| \partial_\sigma^i \partial_\omega^j \partial_\tau^\ell h \right| \leq C \sum_{\ell'' \leq k} \frac{1}{D^{1+\frac{3k-\ell''}{2}}} \left| \partial_{X,\omega,\tau}^{\ell''} C \right|. \tag{6.3.7}$$

Preuve. (a) On utilise les variables (X, ω, τ) et on cherche h sous la forme $h(\sigma, \omega, \tau) = H(X, \omega, \tau)$. L'équation (6.3.5) devient alors $\partial_\tau H + \dfrac{\partial_\sigma w}{D} H = \psi$, qu'on peut écrire, en posant $H = \dfrac{C}{D}$, $\partial_\tau C = D\psi$, soit $C = \int_0^T D\psi\, ds$.

(b) On a alors

$$\begin{aligned}
\partial_\sigma h &= \frac{1}{D} \partial_X C \frac{1}{D} - \frac{C\tau}{D^3} \partial_\sigma^2 w \ , \\
\partial_\omega h &= \frac{1}{D} (\partial_X C \partial_\omega X + \partial_\omega C) - \frac{C}{D^2} (\partial_X D \partial_\omega X + \partial_\omega D) \ , \\
\partial_\tau h &= \frac{1}{D} (\partial_X C \partial_\tau X + \partial_\tau C) - \frac{C}{D^2} (\partial_X D \partial_\tau X + \partial_\tau D),
\end{aligned}$$

ce qui prouve (6.3.6) pour $k = 1$.

(c) La dérivation (en σ par exemple) d'un terme de (6.3.6) produit une somme de la forme

$$* \frac{(\ )^q \partial^{\ell''} C}{D^{2+\ell'+\ell''}} + *q \frac{(\ )^{q-1}}{D^{2+\ell'+\ell''}} \partial^{\ell''} C + * \frac{(\ )^{q+1}}{D^{3+\ell'+\ell''}} \partial^{\ell''} C + * \frac{(\ )^q}{D^{2+\ell'+\ell''}} \partial^{\ell''+1} C \ .$$

Ces termes sont du type voulu, car $q - 1 \geq 2(\ell' + 1) - 3(k + 1 - \ell'')$, $q + 1 \geq 2(\ell' + 2) - 3(k + 1 - \ell'')$ etc.

(d) Puisque $\ell' - \frac{q}{2} \le \frac{3}{2}(k-\ell'')$, $1+\ell'+\ell''-\frac{q}{2} \le 1+\frac{3k}{2}-\frac{\ell''}{2}$, d'où (6.3.7). □

6.3.3. Estimations de F et G. Nous allons maintenant utiliser ces lemmes pour préciser les comportements de F et G définies par (6.1.3) et (6.1.4); on a alors $v = -gF'$, $h = -gG'$, $w = -gF'_A$. On notera $\tilde{X}$ et $\tilde{D}$ les fonctions définies par (6.3.2) et $\tau_*(A,\varepsilon)$ le temps de vie pour ce w là, et on posera, pour $\tau < \tau_*(A,\varepsilon)$

$$
\begin{aligned}
(6.3.8) \qquad & X(\sigma,\omega,\tau) = \tilde{X}(\sigma,\omega,\tau-A),\\
& D(x,\omega,\tau) = 1+(\tau-A)(\partial_\sigma w)(x,\omega).
\end{aligned}
$$

LEMME 6.3.3. *Pour $\tau < A+\tau_*(A,\varepsilon)$, on a les inégalités*

(i) $|F| + |\nabla_{\sigma,\omega,\tau} F| \le C$,

(ii) $|\partial^\alpha_{\sigma,\omega,\tau} F| \le C_\alpha / D(X,\omega,\tau)^{(3k/2)-2}$ *pour* $|\alpha| = k \ge 2$,

(iii) $|G| \le C$, $|\partial^\alpha_{\sigma,\omega,\tau} G| \le C_\alpha / D(X,\omega,\tau)^{3k/2-1/2}$ *pour* $|\alpha| = k \ge 1$.

Preuve. (a) Si $i \ge 1$, $\partial^i_\sigma \partial^j_\omega F = \partial^j_\omega \partial^{i-1}_\sigma \left(\frac{-v}{g}\right)$ et l'on applique (6.3.4). Si $\ell \ge 1$,

$$
\partial_\tau F = \frac{v^2}{2g}, \quad \text{et} \quad (\partial^i_\sigma \partial^j_\omega \partial^{\ell-1}_\tau)\Big(\frac{v^2}{2g}\Big) = \sum_{k'+k''\le k-1} * \partial^{k'} v \, \partial^{k''} v
$$

est majoré, grâce à (6.3.4), par $\dfrac{C}{D^{(3(k'+k''))/2-1}} \le \dfrac{C}{D^{3k/2-2}}$.

(b) Comme

$$
F(\sigma,\omega,\tau) = -\frac{1}{g}\int_M^\sigma v(s,\omega,\tau)\,ds,
$$

on obtient par le changement de variable $x = X(s,\omega,\tau)$, $dx = \dfrac{ds}{D}$,

$$
F(\sigma,\omega,\tau) = -\frac{1}{g}\int_M^{X(\sigma,\omega,\tau)} w(x,\omega)\,D(x,\omega,\tau)\,dx.
$$

Donc

$$
\begin{aligned}
&\partial_\omega(gF(\sigma,\omega,\tau))\\
&\qquad = -(Dw)(X,\omega,\tau)\,\partial_\omega X - \int_M^X \partial_\omega(wD)\,dx = \tau(w\,\partial_\omega w)(X,\omega,\tau) - \int_M^X \ldots dx,
\end{aligned}
$$

et

$$
\partial^2_\omega(gF) = \tau\,\partial_\omega(w\,\partial_\omega w) - \tau^2\,\partial_X(w\,\partial_\omega w)\frac{\partial_\omega w}{D} + \tau\,\partial_\omega(wD)\frac{\partial_\omega w}{D} - \int_M^X \partial^2_\omega(wD)\,dx.
$$

A partir de là, le même raisonnement que pour les fonctions v et h implique (i) et (ii).

(c) En posant, dans (6.1.4), $h = -gG'$, on obtient (6.3.5) avec $\psi(X,\omega,\tau) = -gE(F)$. Rappelons que

$$
E(F) = -\frac{1}{\tau^3}\Big(\frac{1}{4}+\partial^2_\omega\Big)F - \frac{\sigma}{2\tau^2 g} v v' + *v^2 + *v^2 v' + *v\nabla v + *v'\nabla F,
$$

en sorte que $D\psi$ est une fonction régulière de (X,ω,τ), ainsi que C. On déduit donc de (6.3.7) $\left|\partial_\sigma^i\,\partial_\omega^j\,\partial_\tau^\ell h\right| \le \dfrac{C}{D^{1+3k/2}}$.

On a alors, si $i \ge 1$, $\partial_\sigma^i\,\partial_\omega^j G = \partial_\omega^j\,\partial_\sigma^{i-1}\Big(-\dfrac{h}{g}\Big)$; si $\ell \ge 1$,

$$\partial_\tau G = \frac{1}{g}\,v\,h - \frac{1}{g}\int_M^X (D\psi)(x,\omega,\tau)\,dx,$$

d'où

$$\left|\partial_\sigma^i\,\partial_\omega^j\,\partial_\tau^{\ell-1}\,\partial_\tau G\right| \le \sum_{k'+k''\le k-1} \left|\partial^{k'} v\right|\left|\partial^{k''} h\right| + \frac{C}{D^{(3(k-1)/2)-1/2}} \le \frac{C}{D^{3k/2-1}}.$$

Enfin,

$$G(\sigma,\omega,\tau) = -\frac{1}{g}\int_M^\sigma \frac{C}{D}(X,\omega,\tau)\,ds = -\frac{1}{g}\int_M^X C(x,\omega,\tau)\,dx,$$

d'où

$$\partial_\omega(gG) = +\tau\,\partial_\omega w\,\frac{C}{D} - \int_M^X \partial_\omega C\,dx,$$

et (iii). □

6.3.4. Estimations de u_a. Il résulte de ce lemme que, dans la fonction $F+\varepsilon^2 G$, les dérivées de F dominent toujours celles de G pourvu que $\dfrac{\varepsilon^2}{D^{\frac{3}{2}}} \le$ cte. Par ailleurs

$$\partial_i(F(r-t,\omega,\tau)) = F'\omega_i + \partial_\omega F\,O(\varepsilon^2), \qquad \partial_t(F(r-t,\omega,\tau)) = -F' + \frac{\varepsilon}{2\sqrt{t}}\dot{F};$$

pour évaluer les dérivées de u_a, il suffit donc de considérer F et de négliger ω et τ. Cela conduit au lemme suivant.

LEMME 6.3.4. (Dérivées de u_a) *Pour $\tau < A + \tau_*(A,\varepsilon)$, on a les estimations*

(i) $|u_a| + |\nabla_{x,t,\omega} u_a| \le C\varepsilon^2$.

(ii) *Pour $|\alpha| = k \ge 2$,*

$$\left|\partial^\alpha_{x,t,\omega} u_a\right| \le C_\alpha \frac{\varepsilon^2}{D(X,\omega,\tau)^{3k/2-1/2}}.$$

6.3.5. Estimations des dérivées de J_a.

LEMME 6.3.5. *On a, pour $\tau < A + \tau_*(A,\varepsilon)$,*

(i) $|J_a| \le C\dfrac{\varepsilon^8}{D^{\frac{7}{2}}}$,

(ii) $|\nabla_x J_a| + |\partial_\omega J_a| \le C\dfrac{\varepsilon^8}{D^5}$,

(iii) $|\nabla_x^2 J_a| + |\partial_\omega \nabla_x J_a| + |\partial_\omega^2 J_a| \le C\dfrac{\varepsilon^8}{D^{13/2}}$, *où $D = D(X,\omega,\tau)$.*

Preuve. D'après (6.1.3) et les choix de F et G, on a

$$\begin{aligned} J_a = \frac{\varepsilon^6}{\sqrt{rt}}\Big\{ &+ gG'G'' - \frac{1}{\tau^3}\Big(\frac{1}{4}+\partial_\omega^2\Big)G - \frac{\sigma g}{2\tau^2}(F'G''+F''G') - \frac{\dot{G}}{4\tau^2} + \frac{\ddot{G}}{4\tau} \\ &+ g_{ij}^k\Big(\omega_k F' A_{ij} G' + \omega_k G' A_{ij} F' + \omega_i\omega_j F'' A_k G + \omega_i\omega_j G'' A_k F\Big) \\ &+ Q(F+\varepsilon^2 G, F+\varepsilon^2 G) + \varepsilon^2\Big(-\frac{\sigma g}{2\tau^2} G'G'' + g_{ij}^k(\omega_k G' A_{ij} G' + \omega_i\omega_j G'' A_k G)\Big)\Big\}. \end{aligned}$$

Le terme $R(F)$ est contrôlé par les dérivées secondes de F, tandis que le terme le plus singulier entre les accolades est $gG'G''$. On obtient alors les estimations du lemme sans difficulté à partir du lemme (6.3.3). □

6.4. Non dégénérescence et estimations L^2 des erreurs. Sous l'hypothèse (ND) (ou (ND)'), nous précisons maintenant le temps de vie $\tau_*(A,\varepsilon)$ de (6.1.3), ainsi qu'une minoration du dénominateur D des estimations des lemmes (6.3.4)–(6.3.6).

LEMME 6.4.1. *Sous l'hypothèse* (ND), *il existe C tel que, en posant*

$$\tilde\tau = \tilde\tau(\varepsilon) = \tau_*(\varepsilon) - C\varepsilon^{\frac{\kappa}{\kappa-1}} - C\varepsilon^2\,|\log\varepsilon|\,,$$

on ait

(i) $\tau_*(A,\varepsilon) \geq \tilde\tau - A$,

(ii) $\mathrm{cte}\,/(\tilde\tau-\tau) \leq \sup_{\sigma,\omega}\,|gF''(\sigma,\omega,\tau)| \leq \mathrm{cte} + 1/\tilde\tau - \tau$,

(iii) *Il existe une fonction $\sigma_0(\varepsilon)$ et une constante $C_1 > 0$ telles que*

$$D(x,\omega,\tau) \geq \mathrm{cte}\left(\tilde\tau - \tau + C_1(|\omega-\omega_0| + |x-\sigma_0(\varepsilon)|)^\kappa\right).$$

Dans le cas invariant par rotation, et sous l'hypothèse (ND)', (i) et (ii) restent vraies tandis que (iii) est remplacée par

(iii)' $D(x,\tau) \geq \mathrm{cte}\left(\tilde\tau - \tau + C_1\,|x-\sigma_0(\varepsilon)|^\kappa\right)$.

Preuve. (a) D'après l'hypothèse (ND), nous avons

$$-gR''(\sigma,\omega) \geq -gR''(\sigma_0,\omega_0) + C\,d^\kappa\ ,$$

où $d = |\sigma-\sigma_0| + |\omega-\omega_0|$. Cela implique

$$\begin{aligned}&-g(R''+\varepsilon L'')(\sigma,\omega)\\ &\geq -g\Big(R''(\sigma_0,\omega_0)+\varepsilon L''(\sigma_0,\omega_0)\Big) - C'\varepsilon(|\sigma-\sigma_0|+|\omega-\omega_0|) + C\,d^\kappa,\end{aligned}$$

d'où

$$-g(R''+\varepsilon L'')(\sigma,\omega) \geq -g(\quad)(\sigma_0,\omega_0) - C''\varepsilon^{\frac{\kappa}{\kappa-1}} + \frac{C}{2}\,d^\kappa.$$

Les propriétés élémentaires de l'équation de Burger montrent alors que

$$\begin{aligned}-gS''(\sigma,\omega,A) &= \left(A + \frac{1}{-g(R''+\varepsilon L'')(\phi(\sigma,\omega))}\right)^{-1}\\ &\geq \left(A + \frac{1}{-g(\ )(\sigma_0,\omega_0) - C''\varepsilon^{\frac{\kappa}{\kappa-1}} + \frac{C}{2}\,|\phi(\sigma,\omega)-(\sigma_0,\omega_0)|}\right)^{-1},\end{aligned}$$

où $\phi^{-1}(\sigma,\omega)$ correspond au déplacement sur les caractéristiques de l'équation de Burger, à ω fixé, entre $\tau=0$ et $\tau=A$, et $\phi^{-1}(\sigma_0,\omega_0) = (\sigma_0(\varepsilon),\omega_0)$. Comme ϕ est un difféomorphisme et $A < A_0$, $|\phi(\sigma,\omega)-(\sigma_0,\omega_0)| \geq \mathrm{cte}\,\tilde d$, où $\tilde d = |\omega-\omega_0| + |\sigma-\sigma_0(\varepsilon)|$ (ϕ, comme σ_0 dépendent de ε, mais les constantes des estimations n'en dépendent pas).

D'où, pour un $C_1 > 0$,

$$-gS''(\sigma,\omega,A) \geq \left(A + \frac{1}{-g(\ )(\sigma_0,\omega_0)}\right)^{-1} - C\varepsilon^{\frac{\kappa}{\kappa-1}} + C_1\,\tilde d^\kappa\ .$$

Comme

$$\frac{1}{g(\ )(\sigma_0,\omega_0)} = A_0 + \varepsilon A_1 + O(\varepsilon^2),$$

le lemme (6.2.1(i)) implique

$$-gF''_A(\sigma,\omega) \geq \frac{-1}{-A + A_0 + \varepsilon A_1} - C\varepsilon^{\frac{\kappa}{\kappa-1}} - C\varepsilon^2 |\log\varepsilon| + C_1\tilde{d}^\kappa, \tag{6.4.1}$$

avec de plus $-gF''_A(\sigma_0(\varepsilon),\omega_0) = \frac{1}{A-\tau_*} + O(\varepsilon^2 |\log\varepsilon|)$.

De (6.4.1), on déduit que

$$\tau_*(A,\varepsilon) = \inf \frac{1}{gF''_A} \geq \tilde{\tau} - A \ .$$

Comme

$$\sup_{\sigma,\omega} |gF''| \leq \text{cte} + \frac{1}{\tau_*(A,\varepsilon) - (\tau - A)}$$

est vraie pour toute solution d'une équation de Burger, on obtient (ii).

On a enfin

$$D(x,\omega,\tau) = 1 - (\tau - A)gF''_A(x,\omega) \geq \text{cte}\left(\tau_* - \tau - C\varepsilon^{\frac{\kappa}{\kappa-1}} - C\varepsilon^2 |\log\varepsilon| + C_1\tilde{d}^\kappa\right) \ ,$$

d'où (iii). □

Il s'ensuit en particulier que les lemmes 6.3.3–6.3.5 sont vrais pour $\tau < \tilde{\tau}$.

La minoration (iii) de D nous permet d'obtenir des estimations en moyenne meilleures qu'elles ne seraient si on avait $C_1 = 0$.

LEMME 6.4.2. *Il existe $\nu_1 > 0$ tel que, pour $A \leq \tau < \tilde{\tau}$, dans la zone extérieure, on ait*

(i) $|J_a|_0 \leq C\varepsilon^7/(\tilde{\tau} - \tau)^{3-\nu_1}$;

(ii) $|\nabla J_a|_0 + |\partial_\omega J_a|_0 \leq C\varepsilon^7/(\tilde{\tau} - \tau)^{\frac{9}{2}-\nu_1}$;

(iii) $|\nabla^2 J_a|_0 + |\partial_\omega \nabla J_a|_0 + |\partial^2_\omega J_a|_0 \leq C\varepsilon^7/(\tilde{\tau} - \tau)^{6-\nu_1}$.

Preuve. (a) Supposons qu'une fonction $f(\sigma,\omega,\tau) \geq 0$ satisfasse $f(\sigma,\omega,\tau) \leq 1/D(X,\omega,\tau)^\lambda$. On a alors

$$\begin{aligned}\int f^2(\sigma,\omega,\tau)\,d\sigma\,d\omega &\leq \int d\omega \int \frac{d\sigma}{D(X,\omega,\tau)^{2\lambda}} = \int d\omega \int \frac{dx}{D(x,\omega,\tau)^{2\lambda-1}} \\ &\leq \text{cte} \int d\omega \int \frac{dx}{(\tilde{\tau} - \tau + C_1 |x - \sigma_0(\varepsilon)|^\kappa)^{2\lambda-1}}.\end{aligned}$$

En coupant l'intégrale en x en $|x - \sigma_0(\varepsilon)|^\kappa \geq \tilde{\tau} - \tau$ et $\leq \tilde{\tau} - \tau$, il vient

$$\left(\int f^2(\sigma,\omega,\tau)\,d\sigma\,d\omega\right)^{1/2} \leq \frac{\text{cte}}{(\tilde{\tau} - \tau)^{\lambda - \frac{1}{2} - \frac{1}{2\kappa}}}.$$

(b) L'application de la remarque précédente avec $f = |J_a|$ etc., pour les valeurs $\lambda = \frac{7}{2}, 5$ et $\frac{13}{2}$ conduit au résultat, avec $\nu_1 = \frac{1}{2\kappa}$. □

REFERENCES

[1] S. ALINHAC, *Une solution approchée en grand temps des équations d'Euler compressibles axisymétriques en dimension deux*, Comm. Partial Differential Equations, 17 (1992), pp. 447–490.

[2] S. ALINHAC, *Temps de vie et comportement explosif des solutions d'équations d'ondes quasi linéaires en dimension deux*, I, Preprint, Orsay Paris-Sud, 1992; Ann. Sci. Ecole Norm. Sup (4), à paraître.

[3] R. DI PERNA ET A. MAJDA, *The validity of geometrical optics for weak solutions of conservation laws,* Comm. Math. Phys. 98 (1985), pp. 313–347.

[4] G. FRIEDLANDER, *On the radiation field of pulse solutions of the wave equation I, II,* Proc. Roy. Soc. London Ser. A, 269 (1962), pp. 53–65 et 279 (1964), pp. 386–394.

[5] L. HÖRMANDER, *The lifespan of classical solutions of non linear hyperbolic equations,* Mittag-Leffler Report No. 5, 1985.

[6] ———, *The Lifespan of Classical Solutions of Nonliear Hyperbolic Equations*, Lecture Notes in Math., Vol. 1256, Springer-Verlag, New York, Berlin, 1986, pp. 214–280.

[7] S. KLAINERMAN, *Uniform decay estimates and the Lorentz invariance of the classical wave equation,* Comm. Pure Appl. Math., 38 (1985), pp. 321–332.

[8] A. MAJDA, *Compressible fluid flow and systems of conservation laws,* Springer Appl. Math. Sci., Vol. 53, Springer-Verlag, New York, Berlin, 1984.

[9] A. MAJDA ET R. ROSALES, *Resonantly interacting weakly non linear hyperbolic waves I. A single space variable,* Stud. Appl. Math. 71 (1984), pp. 149–179.

[10] W. WASOW, *Asymptotic expansions for ordinary differential equations,* Krieger, New York (1976).

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 566–582, May 1995

003

INSTABILITY OF STATIONARY BUBBLES *

ANNE DE BOUARD†

Abstract. The paper is concerned with the study, by means of a linearized operator, of the instability of stationary solutions of a nonlinear Schrödinger equation with nonvanishing "boundary conditions." We prove that this operator possesses a real positive eigenvalue, and that when they exist, the stationary "bubbles" are always orbitally unstable.

Key words. nonlinear Schrödinger equations, solitary waves, stability

AMS subject classifications. 35B35, 35Q51, 35Q55

1. Introduction. In this work, we study the instability theory of stationary solutions of a nonlinear Schrödinger equation

$$i\frac{\partial \varphi}{\partial t} + \Delta\varphi + F(|\varphi|^2)\varphi = 0 \quad \text{in } \mathbb{R}^n \tag{1}$$

under general assumptions on F. The solutions considered here, unlike the classical standing waves of equation (1), satisfy the particular "boundary condition"

$$\varphi(x,t) \to \sqrt{\rho_0} \quad \text{when } |x| \to +\infty, \tag{2}$$

where ρ_0 is a real positive constant and $F(\rho_0) = 0$. They are the stationary case of a new class of solitary wave solution of equation (1). For example, in the context of a Boson gas with two-body attractive and three-body repulsive interactions, described by the "$\psi^3 - \psi^5$" nonlinear Schrödinger equation

$$i\frac{\partial \psi}{\partial t} + \Delta\psi - \alpha_1\psi + \alpha_3|\psi|^2\psi - \alpha_5|\psi|^4\psi = 0, \tag{3}$$

where $\psi(x,t) \in \mathbb{C}, x \in \mathbb{R}^n, t \in \mathbb{R}_+, n = 1, 2$, or 3, and $\alpha_3, \alpha_5 > 0$, these stationary solutions can be interpreted as "rarefaction bubbles." They represent the nucleation of a stable phase, which is given by the stationary solution $\psi = 0$, into a metastable phase, given by the solution $\psi = \sqrt{\rho_0}$, where 0 and ρ_0 are two minima of the potential V corresponding to equation (3)

$$V(|\psi|^2) = \alpha_1|\psi|^2 - \frac{1}{2}\alpha_3|\psi|^4 + \frac{1}{3}\alpha_5|\psi|^6$$

if $V(0) < V(\rho_0)$ (see [2]).

In space dimension one, some localised solutions traveling with a velocity v, having the form $\psi(x,t) = \varphi(x - vt)$ and corresponding to nonstationary "bubbles" have also been found (see [1]). The "boundary condition" is then

$$\lim_{x \to \pm\infty} \psi(x,t) - \sqrt{\rho_0}e^{\mp i\mu},$$

* Received by the editors September 14, 1992; accepted for publication (in revised form) August 26, 1993.

† Laboratoire d'Analyse Numérique, Université Paris Sud, Bât. 425, 91405 Orsay, Cedex, France.

where μ is a real number depending on the velocity v, with $\mu = 0$ when $v = 0$.

These solutions seem also to have interpretations in other fields of physics where equation (3) occurs (see [2]).

Here, we will be concerned with the general equation (1). Our aim in the paper is to give criteria for the existence of stationary solutions having the same behaviour as those described for equation (3), and to show that for any F satisfying those assumptions and for any space dimension n, such solutions are always unstable. This means that if φ is such a solution, then one can find initial data arbitrarily close to φ such that the corresponding solution of equation (1) quits any given neighborhood of φ in a finite time t_0.

In [2] (see also [4]), Barashenkov et al. considered the linearized evolution equation around such a stationary solution and studied the operators L_1 and L_2 (see §3 for a definition of these operators) to conclude to the existence of a time exponentially growing solution for this equation. We will use a more refined analysis of this linearized operator to show our result (see Theorem 4.1), and in particular, we prove that it possesses a real positive eigenvalue.

The nonlinear Schrödinger equation possesses some other localised solutions which have been extensively studied during the last few years: the stationary states. These solutions are written in the form

$$\varphi(x,t) = e^{i\omega t}u_\omega(x)$$

(see for example [8], [11], [13], [18], [19]). Some stability and instability results for such solutions of equation (1) with a pure power nonlinearity, i.e., when $F(|\varphi|^2)\varphi = \lambda|\varphi|^{p-1}\varphi$, $1 < p < \frac{n+2}{n-2}$, $\lambda > 0$, were proved in [5], [8].

A general theory about stability of solitary waves in Hamiltonian systems was introduced by Grillakis, Shatah, and Strauss [13]. This allows one to obtain stability and instability conditions for some stationary states of equation (1) for more general nonlinearities.

This theory does not seem to apply to the solutions we consider here because these solutions do not belong to $L^2(\mathbb{R}^n)$, and because of the "bad" spectral structure of the Hamiltonian operator.

Let us also mention two papers from Grillakis [11], [12] concerning the study of the linearized operator around a stationary state and the analysis of the instability mechanism.

This work is organized as follows. In §2, we define what we call the "stationary bubbles" and give some existence results for such solutions. Section 3 is devoted to the study of the spectrum of the linearized operator and to the proof of the fact that this operator possesses a real positive eigenvalue with maximal real part. In the last section, we use the result of §3 to show the instability of the stationary bubbles (see Theorem 4.1). The appendix contains the proof of a few technical results used in §§3 and 4.

The notations are as follows. Given equation (1) where φ is a complex valued function of $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$, and where $\Delta\varphi = \sum_{j=1}^n (\partial^2\varphi/\partial x_j^2)$, we assume that F is defined and continuously differentiable in $\mathbb{R}_+$, except in the last section, where F is assumed to be more regular. We also assume that there is a real positive constant ρ_0 such that $F(\rho_0) = 0$ and we set $r_0 = \sqrt{\rho_0}$. For any positive integer m, $H^m(\mathbb{R}^n)$ will denote the Sobolev space of real-valued functions defined on $\mathbb{R}^n$ whose partial derivatives up to order m are in $L^2(\mathbb{R}^n)$. $\mathbb{H}^m(\mathbb{R}^n)$ will denote the space of complex valued functions whose real and imaginary parts are both in $H^m(\mathbb{R}^n)$ and will sometimes be

identified with $H^m(\mathbb{R}^n) \times H^m(\mathbb{R}^n)$. If A is a linear operator on a Banach space, then we will represent its spectrum by $\sigma(A)$, and its resolvent set by $\rho(A) = \mathbb{C}\backslash\sigma(A)$.

2. Existence of the stationary bubbles. In this section, we give some existence results of radially symmetric stationary solutions of equation (1) having the behaviour (2) as $|x|$ goes to infinity. More precisely, according to the definition from Barashenkov et al. [2], a stationary bubble is defined as a real-valued function φ of $x \in \mathbb{R}^n$, satisfying

(i) $\varphi(x) = \varphi(r)$ (i.e., φ is radially symmetric),
(ii) $\Delta\varphi + F(\varphi^2)\varphi = 0$ in $\mathbb{R}^n$,
(iii) $0 < \varphi(r) < r_0$ $\forall r \in [0, \infty)$, $\lim_{r\to+\infty} \varphi(r) = r_0$, and
(iv) $\varphi_r(0) = 0$, $\varphi_r(r) > 0$ $\forall r \in (0, +\infty)$.

Because of condition (iii), we will look for solutions φ such that $u = r_0 - \varphi$ belongs to $H^1(\mathbb{R}^n)$. Hence if φ satisfies (ii), then u is a solution of

$$\Delta u - F((r_0 - u)^2)(r_0 - u) = 0 \quad \text{in } \mathbb{R}^n,\ u \in H^1(\mathbb{R}^n). \tag{4}$$

We set

$$V(s) = -\int_{\rho_0}^{s} F(\tau)\, d\tau. \tag{5}$$

Then we may write the energy associated with equation (1) in terms of $u = r_0 - \varphi$:

$$E(u) = \frac{1}{2}\int_{\mathbb{R}^n} |\nabla u|^2\, dx + \frac{1}{2}\int_{\mathbb{R}^n} V(|r_0 - u|^2)\, dx. \tag{6}$$

We begin with the case $n \geq 2$, since the existence results are not exactly the same as in the case $n = 1$.

2.1. The case $n \geq 2$. In this section we assume that

$$F'(\rho_0) < 0, \tag{7}$$

$$\exists \rho_1 \quad \text{such that } 0 \leq \rho_1 < \rho_0 \quad \text{and} \quad V(\rho_1) < 0, \tag{8}$$

where V is defined by (5).

Then the following result holds.

THEOREM 2.1. *Suppose that F satisfies assumptions* (7) *and* (8); *then there exists a real-valued function $\varphi \in C^2(\mathbb{R}^n)$, which is a stationary bubble, i.e., which satisfies conditions* (i)–(iv). *Moreover, under assumptions* (7) *and* (8), *any stationary bubble φ is twice continuously differentiable in $\mathbb{R}^n$ and satisfies the following property*:

(v) $\exists C > 0$, $\exists \delta > 0$ *such that* $\forall \alpha \in \mathbb{N}^n$ *with* $|\alpha| \leq 2$, $|\partial_x^\alpha(\varphi(x) - r_0)| \leq Ce^{-\delta|x|}, \forall x \in \mathbb{R}^n$.

Proof. This theorem is proved by applying Theorem 1 of Berestycki and Lions [7] if $n \geq 3$ or Berestycki, Gallouët, and Kavian [6] if $n = 2$ to equation (4), or more precisely, to the equation:

$$-\Delta u = g(u), \quad u \in H^1(\mathbb{R}^n), \quad u \neq 0, \tag{9}$$

where g is defined by

$$g(s) = \begin{cases} -F((r_0 - s)^2)(r_0 - s) & \text{if } 0 \leq s \leq r_0, \\ 0 & \text{if } s \geq r_0, \\ -g(-s) & \text{if } s \leq 0. \end{cases}$$

Then there is a positive, radially symmetric, decreasing solution u of (9), and since u is positive, and $u < r_0$ by the maximum principle, u is a solution of (4). Hence, $\varphi = r_0 - u$ is a stationary bubble. Lemmas 1 and 2 in §4 of Berestycki and Lions [7] then show that any function φ satisfying (i)–(iv) is twice continuously differentiable and satisfies (v). □

Remark 2.1. Conditions (7) and (8) will be fulfilled if we assume that V has a minimum at ρ_0 with $V''(\rho_0) > 0$ and at least one more minimum at ρ_1, with $0 < \rho_1 < \rho_0$ and $V(\rho_1) < V(\rho_0)$. It was suggested in [2] that such potentials V should lead to the existence of stationary bubbles.

Remark 2.2. Any solution $u \in H^1(\mathbb{R}^n)$ of the problem $-\Delta u = g(u)$ satisfies the Pohozaev identity (see [7])

$$\frac{n-2}{2} \int_{\mathbb{R}^n} |\nabla u|^2 \, dx = n \int_{\mathbb{R}^n} G(u) \, dx, \tag{10}$$

where G is defined by $G(\xi) = \int_0^\xi g(s) \, ds$. This identity enables one to show that assumption (8) is necessary for the existence of a stationary bubble: if $V(\rho) \geq 0$ for $0 \leq \rho \leq \rho_0$, then such a solution cannot exist.

2.2. The case $n = 1$. This case is of course much simpler, and in fact, we have in this case necessary and sufficient conditions for the existence of a stationary bubble. Also, the solution is unique as soon as it exists. The following result holds.

THEOREM 2.2. *A necessary and sufficient condition of the existence of a stationary bubble φ for equation* (1) *is that*

$$\eta_0 = \sup\{\eta, 0 < \eta < \rho_0, V(\eta) = 0\} \text{ exists}, \qquad 0 < \eta_0 < \rho_0 \quad \text{and} \quad F(\eta_0) < 0. \tag{11}$$

When (11) *is satisfied, such a solution is unique. Moreover if $F'(\rho_0) < 0$, then φ satisfies* (v) (*i.e., $\partial^\alpha(\varphi - r_0)$ is exponentially decreasing for all $|\alpha| \leq 2$*).

Proof. This theorem can be proved by applying Theorem 5 of Berestycki and Lions [7] to the problem

$$-u'' = -F((r_0 - u)^2)(r_0 - u) = g(u). \tag{12}$$

If we set $G(\xi) = \int_0^\xi g(s) \, ds$, then one has $G(2r_0) = 0$ and we can easily show that condition (11) is equivalent to

$$\xi_0 = \inf\{\xi > 0, G(\xi) = 0\} \text{ exists},$$
$$\xi_0 > 0 \quad \text{and} \quad g(\xi_0) > 0,$$

which is the necessary and sufficient condition of Theorem 5 of [7]. □

Remark 2.3. Consider the "$\psi^3 - \psi^5$" nonlinear Schrödinger equation (3). This equation can be rewritten in the following form by a rescaling and a change of function (see [1]), at least in a restricted domain of the parameters α_i:

$$i\frac{\partial \varphi}{\partial t} + \Delta\varphi + (|\varphi|^2 - \rho_0)(2A + \rho_0 - 3|\varphi|^2)\varphi = 0.$$

Here we have $V(\rho) = (\rho - \rho_0)^2(\rho - A)$, and condition (11) is satisfied if and only if $0 < A < \rho_0$. The stationary bubble is then given by

$$\varphi(x) = \frac{\sqrt{\rho_0}\cosh(\kappa x)}{\left(\frac{\rho_0}{A} + \sinh^2(\kappa x)\right)^{1/2}},$$

where

$$\kappa = \sqrt{\rho_0(\rho_0 - A)}.$$

3. The linearized operator. This section is devoted to the study of the linearization of equation (1) around a stationary bubble and to the proof of the linearized instability.

In what follows, φ is a (real-valued) stationary bubble satisfying properties (i)–(v) of §2, and F is supposed to satisfy $F'(\rho_0) < 0$.

Since F is in $\mathcal{C}^1(\mathbb{R}^+)$, the function $h : z \mapsto F(|z|^2)z$ is in $\mathcal{C}^1(\mathbb{R}^2)$, i.e., h' is defined by

$$h'(z)w = \lim_{\epsilon \to 0} \frac{1}{\epsilon}[h(z + \epsilon w) - h(z)]$$

for any $z, w \in \mathbb{C}$. Then if $\varphi + u$ is a solution of (1) with $u(x,t) \in \mathbb{C}, u$ satisfies

$$i\frac{\partial u}{\partial t} + \Delta u + h(\varphi + u) - h(\varphi) = 0$$

and we may write the linearized equation for u as

$$i\frac{\partial u}{\partial t} + \Delta u + h'(\varphi)u = 0$$

or

$$i\frac{\partial u}{\partial t} + \Delta u + F(\varphi^2)u + 2\varphi F'(\varphi^2)\mathcal{R}e(\varphi u) = 0.$$

Hence we set

$$\mathcal{U}(x,t) = \begin{pmatrix} u_1(x,t) \\ u_2(x,t) \end{pmatrix} = (\mathcal{R}e\ u(x,t), \mathcal{I}m\ u(x,t))^t \in \mathbb{R}^2,$$

where $\mathcal{R}e\ u$ is the real part of u and $\mathcal{I}m\ u$ is its imaginary part. We then obtain

$$\frac{\partial \mathcal{U}}{\partial t} = A\mathcal{U} \tag{13}$$

with

$$A = \begin{pmatrix} 0 & L_1 \\ -L_2 & 0 \end{pmatrix}, \qquad \begin{cases} L_1 v = -\Delta v + q_1 v, \\ L_2 v = -\Delta v + (q_1 + q_2)v, \end{cases}$$

and

$$\begin{cases} q_1(x) = q_1(r) = -F(\varphi^2(r)), & r = |x|, \\ q_2(x) = q_2(r) = -2\varphi^2(r)F'(\varphi^2(r)), & r = |x|. \end{cases}$$

We then consider the space $\mathbb{L}^2_r(\mathbb{R}^n)$ of $\mathbb{R}^2$-valued radially symmetric functions of $L^2(\mathbb{R}^n) \times L^2(\mathbb{R}^n)$; A is seen as an unbounded operator on $\mathbb{L}^2_r(\mathbb{R}^n)$. Then, since q_1 and q_2 are in $L^2(\mathbb{R}^n) + L^\infty(\mathbb{R}^n)$, L_1 and L_2 are self-adjoint operators with $D(L_1) = D(L_2) = H^2_r(\mathbb{R}^n) = H^2(\mathbb{R}^n) \cap L^2_r(\mathbb{R}^n)$ (see [17]) and A is closed with maximal domain

$$D(A) = \mathbb{H}^2_r(\mathbb{R}^n) = \mathbb{H}^2(\mathbb{R}^n) \cap \mathbb{L}^2_r(\mathbb{R}^n).$$

Because the aim of this section is to study the spectrum of A, we have to complexify this space. Accordingly, we consider

$$\tilde{\mathbb{L}}^2_r(\mathbb{R}^n) = \mathbb{L}^2_r(\mathbb{R}^n) + i\mathbb{L}^2_r(\mathbb{R}^n),$$

which is identified here with the space of $\mathbb{C}^2$-valued, radially symmetric, square integrable functions. Similarly,

$$\tilde{\mathbb{H}}^2_r(\mathbb{R}^n) = \mathbb{H}^2_r(\mathbb{R}^n) + i\mathbb{H}^2_r(\mathbb{R}^n)$$

with a similar identification. $\tilde{\mathbb{L}}^2_r(\mathbb{R}^n)$ is endowed with its natural inner product

$$((\mathcal{U}, \mathcal{V})) = \mathcal{R}e[\langle u_1, v_1\rangle + \langle u_2, v_2\rangle],$$

where

$$\mathcal{U} = (u_1, u_2)^t, \qquad \mathcal{V} = (v_1, v_2)^t \in \tilde{\mathbb{L}}^2_r(\mathbb{R}^n)$$

and

$$\langle u, v\rangle = \int_{\mathbb{R}^n} u\bar{v}\, dx \quad \text{for } u, v \in \mathbb{L}^2_r(\mathbb{R}^n).$$

We also set

$$\langle\langle \mathcal{U}, \mathcal{V}\rangle\rangle = \langle u_1, v_1\rangle + \langle u_2, v_2\rangle \in \mathbb{C}.$$

We are first interested in locating the "essential spectrum" of the linear unbounded operator A; by this we mean the set of all the values of the spectrum which are not discrete eigenvalues of finite multiplicity (see [17, Vol. IV] for a precise definition). This question is studied in the following lemma.

LEMMA 3.1. *$\sigma_e(A) \subset i\mathbb{R}$ where $\sigma_e(A)$ is the essential spectrum of A.*

Proof. We have

$$\lim_{|x|\to+\infty} q_1(x) = 0 \quad \text{and} \quad \lim_{|x|\to+\infty} q_2(x) = -2\rho_0 F'(\rho_0) = c_0 > 0,$$

and q_1, q_2 converge exponentially to their limit. Thus, A is a relatively compact perturbation of

$$A_0 = \begin{pmatrix} 0 & -\Delta \\ \Delta - c_0 & 0 \end{pmatrix}.$$

Let us first show that $\sigma_e(A_0) \subset i\mathbb{R}$. For any complex number λ,

$$(A_0 + \lambda)(A_0 - \lambda) = A_0^2 - \lambda^2 = \begin{pmatrix} \Delta(-\Delta + c_0) - \lambda^2 & 0 \\ 0 & \Delta(-\Delta + c_0) - \lambda^2 \end{pmatrix} \tag{14}$$

and the spectrum of the differential operator with constant coefficients $\Delta(-\Delta + c_0)$ is imbedded in the set of values of its symbol

$$p(\xi) = -|\xi|^2(|\xi|^2 + c_0).$$

Hence $\sigma(\Delta(-\Delta + c_0)) \subset]-\infty, 0]$ and if $\lambda \notin i\mathbb{R}$ then $\lambda^2 \notin \mathbb{R}^-$ and $A_0^2 - \lambda^2$ is invertible, showing that the nullspace of $A_0 - \lambda$ is $\{0\}$.

Moreover, since if $\lambda^2 \notin]-\infty, 0]$, there is an $\alpha > 0$, depending on λ, such that

$$\frac{\|\xi|^2(|\xi|^2 + c_0) + \lambda^2|}{(1 + |\xi|^2)^2} \geq \alpha > 0 \quad \forall \xi \in \mathbb{R}^n,$$

we have for $v \in H^2(\mathbb{R}^n)$

$$|\langle(-\Delta(-\Delta + c_0) + \lambda^2)v, v\rangle| \geq \alpha\|v\|^2_{H^2}.$$

Hence for $\mathcal{U} \in \tilde{\mathbb{H}}^2_r(\mathbb{R}^n)$

$$|\langle\langle (A_0^2 - \lambda^2)\mathcal{U}, \mathcal{U}\rangle\rangle| \geq \alpha \|\mathcal{U}\|^2_{\tilde{\mathbb{H}}^2_r}.$$

This together with (14) and the fact that $A_0^2 - \lambda^2$ is invertible shows that the range of $A_0 - \lambda$ is $\tilde{\mathbb{L}}^2_r(\mathbb{R}^n)$ and that $\lambda \notin \sigma(A_0)$.

Now, even if Weyl's theorem does not exactly apply here because A is not self-adjoint, an adaptation of the proof of this theorem, using for instance Lemma 3, §XIII-4 of [17] (see Lemma A.1 of the appendix for the statement) shows that the essential spectra of A and A_0 are the same. Indeed, it is not difficult to prove that assumption (ii) of Lemma A.1 is satisfied. The proof of Lemma 3.1 is then complete. □

Remark 3.1. If A has an eigenvalue with positive real part, then it is necessarily an isolated point of the spectrum, by the preceding result. Note that the discrete eigenvalues of A are symmetric with respect to the imaginary axis: if $\mathcal{U} = (u_1, u_2)^t$ is an eigenfunction of A corresponding to an eigenvalue λ, then $(\bar{u}_1, \bar{u}_2)^t$ and $(u_1, -u_2)^t$ are eigenfunctions of A corresponding respectively to $\bar{\lambda}$ and $-\lambda$. We shall show now that A does have an eigenvalue with positive real part.

We begin with the proof of a few properties concerning the self-adjoint operators L_1 and L_2.

LEMMA 3.2. *L_1 is a positive operator, i.e., $(L_1 u, u) > 0$ for any $u \in H^1_r(\mathbb{R}^n)$ with $u \neq 0$.*

Proof. Since $L_1 = -\Delta + q_1(r)$ with $\lim_{r\to+\infty} q_1(r) = 0$, we have $\sigma_e(L_1) = [0, +\infty[$. Now, $L_1\varphi = 0$ and $0 < \varphi(0) \leq \varphi(r) < r_0$ for all $r \geq 0$, thus $q_1(r) = \frac{\Delta\varphi}{\varphi}$ and for $u \in H^1_r(\mathbb{R}^n)$

$$\begin{aligned}\langle L_1 u, u\rangle &= \int_{\mathbb{R}^n} |\nabla u|^2\, dx + \int_{\mathbb{R}^n} \frac{\Delta\varphi}{\varphi}|u|^2\, dx \\ &= \sigma_n \int_0^{+\infty} |u_r|^2 r^{n-1}\, dr + \sigma_n \int_0^{+\infty} (r^{n-1}\varphi_r)_r \frac{|u|^2}{\varphi}\, dr,\end{aligned}$$

σ_n being the measure of the unit sphere in $\mathbb{R}^n$. Hence, integrating by parts, we obtain

$$\begin{aligned}\langle L_1 u, u\rangle &= \sigma_n \int_0^{+\infty} |u_r|^2 r^{n-1}\, dr - 2\sigma_n \int_0^{+\infty} \frac{\varphi_r}{\varphi} \mathcal{R}e(u_r \bar{u}) r^{n-1}\, dr \\ &\quad + \sigma_n \int_0^{+\infty} \frac{\varphi_r^2}{\varphi^2}|u|^2 r^{n-1}\, dr \\ &= \sigma_n \int_0^{+\infty} \varphi^2 \left[\frac{|u_r|^2}{\varphi^2} - \frac{2\varphi_r}{\varphi^3}\mathcal{R}e(u_r\bar{u}) + \frac{|u|^2\varphi_r^2}{\varphi^4}\right] r^{n-1}\, dr \\ &= \sigma_n \int_0^{+\infty} \varphi^2 \left|\frac{u_r}{\varphi} - \frac{u\varphi_r}{\varphi^2}\right|^2 r^{n-1}\, dr \\ &= \sigma_n \int_0^{+\infty} \varphi^2 \left|\left(\frac{u}{\varphi}\right)_r\right|^2 r^{n-1}\, dr\end{aligned}$$

and this leads to

$$\langle L_1 u, u\rangle = \int_{\mathbb{R}^n} \varphi^2 \left|\left(\frac{u}{\varphi}\right)_r\right|^2 dx = \left\|\varphi\left(\frac{u}{\varphi}\right)_r\right\|^2_{L^2(\mathbb{R})^n},$$

which proves the lemma. □

LEMMA 3.3. *$\sigma_e(L_2) = [c_0, +\infty[$; L_2 has at least one negative eigenvalue, and only one if $n = 1$.*

Proof. Since $L_2 = -\Delta + q_1(r) + q_2(r)$ and $\lim_{n\to+\infty} q_2(r) = c_0$, we have $\sigma_e(L_2) = [c_0, +\infty[$.

Assume first that $n \geq 2$; then we have

$$L_1\varphi = 0 = -\varphi_{rr} - \frac{n-1}{r}\varphi_r + q_1(r)\varphi.$$

Differentiating this equation with respect to r, we obtain

$$-\varphi_{rrr} - \frac{n-1}{r}\varphi_{rr} + (q_1(r) + q_2(r))\varphi_r + \frac{n-1}{r^2}\varphi_r = 0.$$

This gives

$$L_2\varphi_r = -\frac{n-1}{r^2}\varphi_r.$$

Since $\varphi \in C^2(\mathbb{R}^n)$ and φ_r is exponentially decreasing when $r \to +\infty$, we have $\varphi_r \in H^1(\mathbb{R}^n)$ and $\varphi_r/r \in L^2(\mathbb{R}^n)$. Moreover $\langle L_2\varphi_r, \varphi_r\rangle < 0$, hence by the min–max theorem, L_2 has at least one negative eigenvalue.

Now if $n = 1$, we have $L_2\dot{\varphi}_r = 0$ and φ_r is a generator of the kernel of L_2 in $L^2(\mathbb{R}^n)$, because any solution of the differential equation $L_2u = 0$ on $[0, +\infty[$, independent of φ_r satisfies $u(x) \sim u_0e^{c_0x}$ where $u_0 \in \mathbb{C}$ and thus is not in $L^2(\mathbb{R}^n)$. Since φ vanishes only for $r = 0$, we deduce by the Sturm–Liouville theory that L_2 has a unique negative eigenvalue, corresponding to an always-positive eigenfunction. □

Remark 3.2. $0 \in \sigma_e(L_1)$, L_1 is self-adjoint, and 0 is not an eigenvalue for L_1, consequently, the range of L_1 is dense in $L^2(\mathbb{R}^n)$ and we can define L_1^{-1} as an unbounded operator with dense domain in $L^2(\mathbb{R}^n)$, as was noted in [2]. Assume now that $\mathcal{U} = (u_1, u_2)^t$ is an eigenfunction for A corresponding to a real eigenvalue λ, i.e.,

$$\begin{cases} L_1u_2 = \lambda u_1, \\ -L_2u_1 = \lambda u_2. \end{cases}$$

Then $u_1 \in D(L_1^{-1}), u_2 = \lambda L_1^{-1}u_1$, and $(L_2 + \lambda^2 L_1^{-1})u_1 = 0$, which tells us that u_1 is in the nullspace of $L_2 + \lambda^2 L_1^{-1}$. Conversely, if we prove that there is a positive γ such that $L_2 + \gamma L_1^{-1}$ has a nontrivial nullspace $N(L_2 + \gamma L_1^{-1})$ and $D(L_2 + \gamma L_1^{-1}) = D(L_2) \cap D(L_1^{-1})$, then taking $u_1 \in N(L_2 + \gamma L_1^{-1})$ and $u_2 = \sqrt{\gamma} L_1^{-1}u_1$, we obtain an eigenfunction $(u_1, u_2)^t$ of A for the positive eigenvalue $\sqrt{\gamma}$.

Thus we have to study the operator $L_2 + \lambda^2 L_1^{-1}$ for $\lambda \in \mathbb{R}^*$.

LEMMA 3.4. *For any $\lambda \in \mathbb{R}^*$, the operator $L_2 + \lambda^2 L_1^{-1}$ satisfies the following properties:*

(i) *$L_2 + \lambda^2 L_1^{-1}$ is self-adjoint on its maximal domain, which is $D(L_2) \cap D(L_1^{-1})$.*

(ii) *$\sigma_e(L_2 + \lambda^2 L_1^{-1}) \subset [c_0, +\infty[$.*

Proof. (i) Since $L_2 + \lambda^2 L_1^{-1} = L_1 + \lambda^2 L_1^{-1} + q_2$ and $u \mapsto q_2u$ is a bounded self-adjoint operator on $L^2_r(\mathbb{R}^n)$, it suffices to prove that $D(L_1 + \lambda^2 L_1^{-1}) = D(L_1) \cap D(L_1^{-1})$, and that $L_1 + \lambda^2 L_1^{-1}$ is self-adjoint on $D(L_1) \cap D(L_1^{-1})$. The result is then obtained by applying the Kato–Rellich theorem (see [17, Vol. IV]).

Now, since L_1 is self-adjoint, positive, and since 0 is not an eigenvalue of L_1, the spectral theorem gives

$$L_1 + \lambda^2 L_1^{-1} = \int_{-\infty}^{+\infty} \left(\mu + \frac{\lambda^2}{\mu}\right) dP_\mu.$$

Thus $L_1 + \lambda^2 L_1^{-1}$ is self-adjoint on

$$D(L_1 + \lambda^2 L_1^{-1}) = \left\{ u \in L^2(\mathbb{R}^n), \int_{-\infty}^{+\infty} \left(\mu + \frac{\lambda^2}{\mu}\right)^2 d(P_\mu u, u) < +\infty \right\}.$$

But if $u \in D(L_1 + \lambda^2 L_1^{-1})$ then

$$\int_{-\infty}^{+\infty} \mu^2 \, d(P_\mu u, u) < +\infty$$

and

$$\int_{-\infty}^{+\infty} \frac{1}{\mu^2} \, d(P_\mu u, u) < +\infty.$$

Hence $u \in D(L_1) \cap D(L_1^{-1})$. This proves (i).

To prove (ii) let us first show that

$$\sigma_e(L_2 + \lambda^2 L_1^{-1}) = \sigma_e(-\Delta + c_0 + \lambda^2 L_1^{-1}).$$

Using Weyl's theorem, we only have to prove that $u \mapsto (q_1 + q_2 - c_0)u$ is relatively compact with respect to $-\Delta + c_0 + \lambda^2 L_1^{-1}$. Consider $D(-\Delta + c_0 + \lambda^2 L_1^{-1}) = H_r^2(\mathbb{R}^n) \cap D(L_1^{-1})$, endowed with the graph norm

$$|||u|||^2 = \|(-\Delta + \lambda^2 L_1^{-1})u\|_{L^2}^2 + \|u\|_{L^2}^2.$$

Then we have

$$\begin{aligned} |||u|||^2 &= \|\Delta u\|_{L^2}^2 + \lambda^4 \|L_1^{-1} u\|_{L^2}^2 + 2\lambda^2 \mathcal{R}e\langle -\Delta u, L_1^{-1} u\rangle + \|u\|_{L^2}^2 \\ &= \|\Delta u\|_{L^2}^2 + \lambda^4 \|L_1^{-1} u\|_{L^2}^2 + 2\lambda^2 \mathcal{R}e\langle L_1 u, L_1^{-1} u\rangle - 2\lambda^2 \mathcal{R}e\langle q_1 u, L_1^{-1} u\rangle + \|u\|_{L^2}^2 \\ &\geq \|\Delta u\|_{L^2}^2 + \lambda^4 \|L_1^{-1} u\|_{L^2}^2 + 2\lambda^2 \|u\|_{L^2}^2 + \|u\|_{L^2}^2 - \lambda^4 \|L_1^{-1} u\|_{L^2}^2 - \alpha^2 \|u\|_{L^2}^2 \end{aligned}$$

with $\alpha = \sup_{r \geq 0} q_1(r)$. Hence if $u \in D(L_1^{-1}) \cap H_r^2$,

$$\|\Delta u\|_{L^2}^2 + (2\lambda^2 + 1)\|u\|_{L^2}^2 \leq |||u|||^2 + \alpha^2 \|u\|_{L^2}^2$$

and $D(-\Delta + c_0 + \lambda^2 L_1^{-1})$ is continuously imbedded in $H_r^2(\mathbb{R}^n)$. Now, since $u \mapsto (q_1 + q_2 - c_0)u$ is compact from $H_r^2(\mathbb{R}^n)$ into $L_r^2(\mathbb{R}^n)$, we obtain by Weyl's essential spectrum theorem that

$$\sigma_e(L_2 + \lambda^2 L_1^{-1}) = \sigma_e(-\Delta + c_0 + \lambda^2 L_1^{-1}).$$

On the other hand, we have

$$\inf_{\substack{u \in D(-\Delta + c_0 + \lambda^2 L_1^{-1}) \\ \|u\|_{L^2} = 1}} \langle (-\Delta + c_0 + \lambda^2 L_1^{-1})u, u\rangle \geq c_0,$$

and hence

$$\sigma_e(L_2 + \lambda^2 L_1^{-1}) \subset \sigma_e(-\Delta + c_0 + \lambda^2 L_1^{-1}) \subset [c_0, +\infty[,$$

which proves (ii). □

Now we are able to prove the following theorem about the existence of a real eigenvalue for A.

THEOREM 3.1. *The operator*

$$A = \begin{pmatrix} 0 & L_1 \\ -L_2 & 0 \end{pmatrix}$$

defined on $\tilde{\mathbb{L}}^2_r(\mathbb{R}^n)$ *has a real positive eigenvalue with maximal real part, given by*

$$\lambda_{max} = \left[- \inf_{u \in H^1_r(\mathbb{R}^n) \cap D(L_1^{-1})} \frac{\langle L_2 u, u \rangle}{\langle L_1^{-1} u, u \rangle} \right]^{1/2}.$$

Proof. According to Lemma 3.4, L_2 has a negative eigenvalue. On the other hand, $D(L_1^{-1}) = R(L_1)$ is dense in $L^2_r(\mathbb{R}^n)$. In fact it can be proved that $D(L_1^{-1}) \cap H^1_r(\mathbb{R}^n)$ is dense in $H^1_r(\mathbb{R}^n)$ (see [9] for the details). Hence, there is a $u_0 \in D(L_1^{-1}) \cap H^1_r(\mathbb{R}^n)$ such that $\langle u_0, L_2 u_0 \rangle < 0$. Since L_1^{-1} is positive, we deduce that

$$-\infty \leq \inf_{\substack{u \neq 0 \\ u \in D(L_1^{-1}) \cap H^1_r(\mathbb{R}^n)}} \frac{\langle u, L_2 u \rangle}{\langle u, L_1^{-1} u \rangle} < 0.$$

Let us show that this expression is finite. If $u \in H^1_r(\mathbb{R}^n) \cap D(L_1^{-1})$, then with $L_1^{1/2}$ and $L_1^{-1/2}$ as defined by the spectral theorem, we have

$$\begin{aligned} \langle L_2 u, u \rangle &= \langle L_1 u, u \rangle + \langle q_2 u, u \rangle \\ &\geq \langle L_1 u, u \rangle - \beta \|u\|^2_{L_2(\mathbb{R}^n)}, \qquad \beta = \sup_{r \geq 0} q_2(r) \\ &\geq \langle L_1 u, u \rangle - \beta \langle L_1^{1/2} u, L_1^{-1/2} u \rangle \\ &\geq \langle L_1 u, u \rangle - \beta \langle L_1 u, u \rangle^{1/2} \langle L_1^{-1} u, u \rangle^{1/2} \\ &\geq -\frac{\beta^2}{4} \langle L_1^{-1} u, u \rangle, \end{aligned}$$

where we have used Young's inequality in the last estimate. Hence,

$$\inf_{\substack{u \neq 0 \\ u \in D(L_1^{-1}) \cap H^1_r(\mathbb{R}^n)}} \frac{\langle u, L_2 u \rangle}{\langle u, L_1^{-1} u \rangle} \geq -\frac{\beta^2}{4} > -\infty.$$

Now, let $u_n \in D(L_1^{-1}) \cap H^1_r(\mathbb{R}^n)$, with $\|u_n\|^2_{L^2} = 1$ and

$$\lim_{n \to +\infty} \frac{\langle u_n, L_2 u_n \rangle}{\langle u_n, L_1^{-1} u_n \rangle} = \inf_{\substack{u \neq 0 \\ u \in D(L_1^{-1}) \cap H^1_r(\mathbb{R}^n)}} \frac{\langle u, L_2 u \rangle}{\langle u, L_1^{-1} u \rangle} = -\gamma.$$

Then $\langle u_n, L_1^{-1} u_n \rangle$ is bounded, since if it was not, $\langle u_n, L_2 u_n \rangle$ would not be bounded either. But $\langle u_n, L_2 u_n \rangle$ is negative for large n and

$$\langle u_n, L_2 u_n \rangle = \langle u_n, L_1 u_n \rangle + \langle u_n, q_2 u_n \rangle \geq -\beta \|u_n\|^2_{L_2} = -\beta,$$

hence $\langle u_n, L_2 u_n \rangle$ is bounded.

This proves that

$$\lim_{n\to+\infty}\langle L_2u_n+\gamma L_1^{-1}u_n,u_n\rangle=0.$$

Now, for all u in $H_r^1\cap D(L_1^{-1})$, $\langle(L_2+\gamma L_1^{-1})u,u\rangle\geq 0$ and 0 is exactly the infimum of the spectrum of $L_2+\gamma L_1^{-1}$. Moreover, since from Lemma 3.4, $\sigma_e(L_2+\gamma L_1^{-1})\subset[c_0,+\infty[$, 0 is necessarily an eigenvalue for $L_2+\gamma L_1^{-1}$, i.e., $N(L_2+\gamma L_1^{-1})\neq 0$ and using the argument in Remark 3.2, this implies that $\sqrt{\gamma}$ is an eigenvalue for A.

If $\lambda^2>\gamma$ and $u\in H_r^2\cap D(L_1^{-1}), u\neq 0$, then

$$\begin{aligned}\langle(L_2+\lambda^2L_1^{-1})u,u\rangle&\geq\langle(L_2+\gamma L_1^{-1})u,u\rangle+(\lambda^2-\gamma)\langle L_1^{-1}u,u\rangle\\&\geq(\lambda^2-\gamma)\langle L_1^{-1}u,u\rangle>0.\end{aligned}$$

Hence we have $N(L_2+\lambda^2L_1^{-1})=0,\forall\lambda>\sqrt{\gamma}$, and $\sqrt{\gamma}$ is a maximal eigenvalue for A. □

Remark 3.3. If z is a (not necessarily real) eigenvalue for A, and if $(u_1,u_2)^t$ is a corresponding eigenvector, then

$$z^2=-\frac{\langle L_2u_1,u_1\rangle}{\langle L_1^{-1}u_1,u_1\rangle}$$

and hence $z^2\in\mathbb{R}$. This shows that $\sqrt{\gamma}$ is of maximal real part among all the eigenvalues of A.

Remark 3.4. For the case of the usual stationary states, which are solutions of equation (1.1) having the form $\varphi(x,t)=e^{i\omega t}u_\omega(x)$ with u_ω satisfying the equation

$$\Delta u_\omega-\omega u_\omega+F(|u_\omega|^2)u_\omega=0,\qquad u_\omega\in H^1(\mathbb{R}^n),$$

linearizing around one of those real-valued solutions leads to the linearized operator

$$A_\omega=\begin{pmatrix}0&L_1^\omega\\-L_2^\omega&0\end{pmatrix}$$

with

$$\begin{cases}L_1^\omega=-\Delta-F(u_\omega^2)+\omega,\\L_2^\omega=-\Delta-F(u_\omega^2)-2u_\omega^2F'(u_\omega^2)+\omega.\end{cases}$$

Although A_ω has a form similar to our operator A, the main difference is that 0 is an eigenvalue for L_1^ω since $u_\omega\in L^2(\mathbb{R}^n)$. If we define P as the orthogonal projection on $[N(L_1^\omega)]^\perp$, the existence of real eigenvalues for A_ω can be proved by studying the kernel of $PL_2^\omega P+\lambda^2(L_1^\omega)^{-1}$ defined on $[N(L_1^\omega)]^\perp$. Now, in our proof of Theorem 3.1, the fact that L_2 has at least one negative eigenvalue was essential. But here, even if L_2^ω always has a negative eigenvalue, it may happen that $PL_2^\omega P$ is a positive operator and hence that $N(PL_2^\omega P+\lambda^2(L_1^\omega)^{-1})=\{0\}$ for all $\lambda\in\mathbb{R}$. This shows that in the case of stationary states, unlike the case of stationary bubbles, a more refined study of the operator $PL_2^\omega P+\lambda^2(L_1^\omega)^{-1}$ has to be done to draw conclusions about the stability or instability of the solutions (see the works of Grillakis [11] and [12] for some results in this direction).

4. Instability of the stationary bubbles. In this section, we prove the main result of the paper (Theorem 4.1), which is the instability of the stationary bubbles, using the results of the preceding section concerning the linearized operator. But to

prove such a result, some regularity has to be assumed for F, and we will also need some regularity about the bubbles we consider.

We assume in what follows that (7) and (8) are satisfied, and thus by Theorems 2.1 and 2.2, any stationary bubble satisfies property (v) of §2.

The following proposition shows that the stationary bubbles are regular when F is.

PROPOSITION 4.1. *Assume that F is in $\mathcal{C}^m(\mathbb{R}^+)$ with m a positive integer. Let φ be a stationary bubble; then $r_0 - \varphi \in H^{m+2}(\mathbb{R}^n)$.*

Proof. We set $u = r_0 - \varphi$. Since φ satisfy properties (i)–(v) of §2, we have $u \in H^2(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$, and u satisfies

$$-\Delta u = g(u)$$

with $g(s) = -F((r_0 - s)^2)(r_0 - s)$, i.e., $g \in \mathcal{C}^m(\mathbb{R}^2)$ and $g(0) = 0$. The result then follows from Moser's composition inequality ([15, I-2]).

In what follows we assume that $F \in \mathcal{C}^{m+2}(\mathbb{R}^n)$, with m an integer larger than $\frac{n}{2}$. We consider the space of functions $u \in \mathbb{H}^m(\mathbb{R}^n)$ defined in §1, this space is identified with $H^m(\mathbb{R}^n) \times H^m(\mathbb{R}^n)$.

We give a very classical result concerning the existence of solutions φ for the evolution equation (1), such that $r_0 - \varphi \in H^m(\mathbb{R}^n)$. □

PROPOSITION 4.2. *Let $u_0 \in \mathbb{H}^m(\mathbb{R}^m)$. Then there exist $T_+, T_- \in]0, +\infty]$ depending only on $\|u_0\|_{\mathbb{H}^m}$ such that the equation*

$$i\frac{\partial u}{\partial t} + \Delta u - F(|r_0 - u|^2)(r_0 - u) = 0 \tag{15}$$

has a unique solution $u \in \mathcal{C}^0(]T_-, T_+[, \mathbb{H}^m(\mathbb{R}^n)) \cap \mathcal{C}^1(]T_-, T_+[, \mathbb{H}^{m-2}(\mathbb{R}^n))$ satisfying $u(0) = u_0$.

Proof. The proof is obvious since $i\Delta$ generates a unitary group in $\mathbb{H}^m(\mathbb{R}^n)$ and $v \mapsto F(|r_0 - v|^2)(r_0 - v)$ is locally Lipschitz on $\mathbb{H}^m(\mathbb{R}^n)$ for $m > \frac{n}{2}$ (see for example the proof of Proposition A.1). □

The following theorem is the main result of the paper.

THEOREM 4.1. *Let φ be a stationary bubble (satisfying properties* (i)–(v) *of* §2). *Then φ is unstable in the following sense. $\exists \varepsilon > 0, \forall \delta > 0, \exists u_0 \in \mathbb{H}^m(\mathbb{R}^n)$ satisfying $\|u_0\|_{\mathbb{H}^m} < \delta$ and such that if $u(t) \in \mathbb{H}^m(\mathbb{R}^n)$ and $v(t) = u(t) + \varphi$ is the solution of equation* (1) *with $v(0) = u_0 + \varphi$, then $\exists t_0 > 0$ such that $\|u(t_0)\|_{\mathbb{H}^m} > \varepsilon$.*

Proof. Let φ be a stationary bubble; if $v = \varphi + u$ is a solution of equation (1), then u is a solution of

$$i\frac{\partial u}{\partial t} + \Delta u + h'(\varphi)u + [h(\varphi + u) - h(\varphi) - h'(\varphi)u] = 0,$$

where $h(z) = F(|z|^2)z$. From Proposition A.1, we get

$$\|h(\varphi + u) - h(\varphi) - h'(\varphi)u\|_{\mathbb{H}^m} = O(\|u\|^2_{\mathbb{H}^m}) \quad \text{as } \|u\|_{\mathbb{H}^m} \to 0.$$

Hence u satisfies an equation of the form

$$\frac{\partial \mathcal{U}}{\partial t} = A\mathcal{U} + f(\mathcal{U}), \tag{16}$$
$$\mathcal{U} = (u_1, u_2)^t = (\mathcal{R}e\ u, \mathcal{I}m\ u)^t,$$

where A is the operator defined in §3 and

$$\|f(\mathcal{U})\|_{\mathbb{H}^m} = O(\|\mathcal{U}\|^2_{\mathbb{H}^m}) \quad \text{as } \|\mathcal{U}\|^2_{\mathbb{H}^m} \to 0.$$

It follows from Proposition 4.1 that $r_0 - \varphi \in H^m(\mathbb{R}^n)$. Therefore, $q_1 = F(\varphi^2)$ and $q_2 - c_0 = 2\varphi^2 F'(\varphi^2) - c_0$ are in $H^m(\mathbb{R}^n)$, and for any $\mathcal{U} = (u_1, u_2)^t \in \mathbb{H}^m$, the following holds:

$$\begin{aligned}((A\mathcal{U},\mathcal{U})) &= (L_1 u_2, u_1)_m - (L_2 u_1, u_2)_m \\ &= (q_1 u_2, u_1)_m - ((q_1 + q_2)u_1, u_2)_m \\ &\le (2\|q_1\|_{H^m} + \|q_2\|_{L^\infty} + \|q_2 - c_0\|_{H^m})\|\mathcal{U}\|^2_{\mathbb{H}^m},\end{aligned}$$

where we still denote by $((\cdot,\cdot))$ the inner product of the real Hilbert space $\mathbb{H}^m(\mathbb{R}^n)$, and by $(\cdot,\cdot)_m$ the inner product of $H^m(\mathbb{R}^n)$.

This shows that A generates a $\mathcal{C}^0$ semigroup $S(t)$ in $\mathbb{H}^m(\mathbb{R}^n)$ satisfying

$$\|S(t)\|_{\mathcal{L}(\mathbb{H}^m)} \le Me^{\omega t}, \tag{17}$$

where $e^{\omega t}$ is the spectral radius of $S(t)$ and

$$\omega = \inf_{t>0}\left\{\frac{\ln\|S(t)\|}{t}\right\} \ge \sqrt{\lambda_{\max}} > 0.$$

Then Proposition A.2 applies with $X = \mathbb{H}^m(\mathbb{R}^n)$ and the instability of the fixed point 0 in equation (16) follows, which means that there is a positive ε such that for any positive δ, one can find $\mathcal{U}_0 \in \mathbb{H}^m(\mathbb{R}^n)$ with $\|\mathcal{U}_0\|_{\mathbb{H}^m} < \delta$ such that if $\mathcal{U}(t)$ is the solution of equation (16) with $\mathcal{U}(0) = \mathcal{U}_0$, then there is a positive t_0 for which $\|\mathcal{U}(t_0)\|_{\mathbb{H}^m} > \varepsilon$. This proves Theorem 4.1. □

Remark 4.1. Although the linearized operator A possesses an eigenfunction $u_0 \in H^m(\mathbb{R}^n)$ corresponding to its maximal eigenvalue $\lambda_{\max} > 0$, it remains unclear whether this eigenfunction provides a nonlinearly unstable solution $u(t)$. The difficulty lies in the fact that the real number ω in estimate (17) may be greater than $\lambda_{\max}$. The existence of a positive eigenvalue $\lambda_{\max}$ only allows us to say that ω is positive.

Remark 4.2. When a family of traveling waves $\varphi_v(x - vt)$ is considered, with $v \mapsto \varphi_v$ continuous with values in X, then one has to study the stability modulo translations, because if $|v_1 - v_2|$ is small then $|\varphi_{v_1} - \varphi_{v_2}|_X$ is also small, but one cannot expect $\sup_{t\in\mathbb{R}} |\varphi_{v_1}(\cdot - v_1 t) - \varphi_{v_2}(\cdot - v_2 t)|_X$ to be small since the two traveling waves are propagating with different velocities. In the case of the bubbles, in dimension $n = 1$, the family of traveling waves is not continuous with values in $H^1(\mathbb{R}^n)$ and one has moreover if $v_1 \ne v_2, \lim_{|x|\to\infty} \varphi_{v_1}(x) \ne \lim_{|x|\to\infty} \varphi_{v_2}(x)$, so that $\varphi_{v_1} - \varphi_{v_2}$ is not in $H^1(\mathbb{R}^n)$. In addition, a paper from Barashenkov and Panova [3] tends to show by numerical computations that the (one-dimensional) cubic-quintic traveling bubbles, which seem to be unstable for low velocities and to become stable after some critical velocity v_c, present a real stability for $v > v_c$ and not only a stability modulo translations. Therefore it is not obvious whether this form of instability has to be considered here. However, the remark after Theorem 2 in [14] allows us to say that the stationary bubbles are indeed unstable modulo translations, since if φ is a stationary bubble, then the family $\{\varphi(\cdot - \alpha) - \varphi, \alpha \in \mathbb{R}^n\}$ is a C^1 family with values in $H^1(\mathbb{R}^n)$.

Appendix. The first lemma, which is a well-known adaptation of Weyl's essential spectrum theorem to the case of closed unbounded operators, follows for instance from Lemmas 2 and 3, §XIII-4 in [17].

LEMMA A.1. *Let A and B be closed unbounded operators with dense domain on a Banach space. We assume that*

(i) *$\sigma(A)$ has an empty interior set in $\mathbb{C}$,*

(ii) *each connected component of $\mathbb{C}\backslash\sigma(A)$ contains a point in $\rho(B)$, and*

(iii) *there is a λ_0 in $\rho(B)\cap\rho(A)$ such that $(A-\lambda_0)^{-1}-(B-\lambda_0)^{-1}$ is a compact operator.*

Then $\sigma_e(A) = \sigma_e(B)$.

Proof. Using Lemma 2, §XIII-4 in [17], one can connect $\sigma_e(A)$ and $\sigma_e((A-\lambda_0)^{-1})$; then [17, Lem. 3, §XIII-4] applies to $(A-\lambda_0)^{-1}$ and $(B-\lambda_0)^{-1}$, which are bounded operators. □

Now, we prove a proposition which allows us to linearize equation (1) in $\mathbb{H}^m(\mathbb{R}^n)$ if $m > \frac{n}{2}$ and which is used in the proof of Theorem 4.1.

PROPOSITION A.1. *Let m be an integer greater than $\frac{n}{2}$, and let $g : \mathbb{R}^2 \to \mathbb{R}^2$ be a function of class C^{m+2} on $\mathbb{R}^2$. Then for any $v \in \mathbb{H}^m(\mathbb{R}^n) = H^m(\mathbb{R}^n) \times H^m(\mathbb{R}^n)$ there are positive constants a and b, depending only on $\|v\|_{\mathbb{H}^m}$ such that, for any $u \in \mathbb{H}^m(\mathbb{R}^n)$ with $\|u\|_{\mathbb{H}^m} \le a$,*

$$\|g(u+v) - g(v) - g'(v)u\|_{\mathbb{H}^m} \le b\|u\|^2_{\mathbb{H}^m}.$$

Proof. Let $u = (u_1, u_2)^t, v = (v_1, v_2)^t \in \mathbb{H}^m(\mathbb{R}^n)$. Then

$$\begin{aligned} g(u+v) - g(v) - g'(v)u &= \int_0^1 (1-t)g''(v+tu)(u,u)\,dt \\ &= \sum_{i,j=1}^{2} u_i u_j \int_0^1 (1-t)\frac{\partial^2 g}{\partial y_i \partial y_j}(v+tu)\,dt. \end{aligned}$$

Hence

$$\begin{aligned} &\|g(v+u) - g(v) - g'(v)u\|_{\mathbb{H}^m} \\ &\quad \le \|u\|^2_{\mathbb{H}^m}\left(C + C'\int_0^1 (1-t)\left\|\frac{\partial^2 g}{\partial y_i \partial y_j}(v+tu) - \frac{\partial^2 g}{\partial y_i \partial y_j}(0)\right\|_{\mathbb{H}^m} dt\right). \end{aligned}$$

Since $\partial_{y_i}\partial_{y_j} g_k \in C^m(\mathbb{R}^2)$ for $i, j, k \in \{1, 2\}$, it follows that

$$\left\|\frac{\partial^2 g}{\partial y_i \partial y_j}(v) - \frac{\partial^2 g}{\partial y_i \partial y_j}(0)\right\|_{\mathbb{H}^m} \le C(\|v\|_{L^\infty})(1 + \|v\|_{\mathbb{H}^m}),$$

as in the proof of Proposition 4.1.

The last term is bounded with $\|v\|_{\mathbb{H}^m}$ and this proves Proposition A.1. □

The following proposition, which makes the connection between linearized and nonlinear instability, follows from a theorem due to Henry, Perez, and Wreszinski [14]; we recall it here for the sake of completeness, but in a slightly weaker form, which is in fact sufficient in our case.

PROPOSITION A.2. *Let X be a Banach space, A a linear unbounded operator on X, and consider the equation*

$$\frac{\partial \mathcal{U}}{\partial t} = A\mathcal{U} + f(\mathcal{U}), \tag{18}$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ satisfies

$$\|f(\mathcal{U})\|_X = O(\|\mathcal{U}\|^2_X) \quad \text{as } \|\mathcal{U}\|_X \to 0.$$

Assume that A generates a $\mathcal{C}^0$ semigroup $S(t)$ on X satisfying $\|S(t)\|_{\mathcal{L}(X)} \leq Me^{\omega t}$ with

$$\omega = \inf_{t>0} \left\{ \frac{\|S(t)\|_{\mathcal{L}(X)}}{t} \right\} > 0.$$

Then 0 is an unstable fixed point of equation (18), i.e., $\exists \varepsilon > 0, \forall \delta > 0, \exists \mathcal{U}_0 \in X$ with $\|\mathcal{U}_0\|_X < \delta$ and $\exists t_0 > 0$ such that $\|\mathcal{U}(t_0)\|_X > \varepsilon$ where $\mathcal{U}(t)$ is the solution of (18) with $\mathcal{U}(0) = \mathcal{U}_0$.

Proof. We denote by $T(t)$ the nonlinear semigroup associated with equation (18), i.e., $T(t)\mathcal{U}_0 = \mathcal{U}(t)$ where $\mathcal{U}(t)$ is the solution of (18) with $\mathcal{U}(0) = \mathcal{U}_0$. Also, we denote by $\|\cdot\|$ the norm on the Banach space X.

Since $f(\mathcal{U}) = O(\|\mathcal{U}\|^2)$ as $\|\mathcal{U}\| \to 0$ there are constants $a_1 > 0$ and $c > 0$ such that

$$\|\mathcal{U}\| < a_1 \Rightarrow \|f(\mathcal{U})\| \leq c\|\mathcal{U}\|^2.$$

Assume that 0 is a stable fixed point of $T(t)$, i.e., $\forall \varepsilon > 0, \exists \alpha > 0, \|\mathcal{U}_0\| < \alpha \Rightarrow \|T(t)\mathcal{U}_0\| < \varepsilon$ for all positive t. First, let us take $\varepsilon = a_1$; then $\exists a_0 > 0$ such that $\|\mathcal{U}_0\| < a_0 \Rightarrow \|\mathcal{U}(t)\| < a_1, \forall t \geq 0$, where $\mathcal{U}(t) = T(t)\mathcal{U}_0$.

But we have

$$T(t)\mathcal{U}_0 = S(t)\mathcal{U}_0 + \int_0^t S(t-r) f(\mathcal{U}(r))\, dr. \tag{19}$$

Thus, we have for $t \in [0, t_0]$ and $\|\mathcal{U}_0\| < a_0$

$$\begin{aligned} \|\mathcal{U}(t)\| &\leq Me^{\omega t}\|\mathcal{U}_0\| + \int_0^t \|S(t-r)\|_{\mathcal{L}(X)} \|f(\mathcal{U}(r))\|\, dr \\ &\leq Me^{\omega t}\|\mathcal{U}_0\| + a_1 MC \int_0^t e^{\omega(t-r)} \|\mathcal{U}(r)\|\, dr, \end{aligned}$$

and using the Gronwall lemma,

$$\|\mathcal{U}(t)\| \leq e^{\omega t_0}(M + e^{a_1 MCt_0})\|\mathcal{U}_0\| \tag{20}$$

as soon as $t \in [0, t_0]$ and $\|\mathcal{U}_0\| < a_0$.

Then using equation (19) again, we obtain for $t_0 > 0$ and $\|\mathcal{U}_0\| < a_0$

$$\begin{aligned} \|T(t)\mathcal{U}_0 - S(t)\mathcal{U}_0\| &\leq \int_0^{t_0} \|S(t_0 - r)\| \|f(\mathcal{U}(r))\|\, dr \\ &\leq MC \int_0^{t_0} e^{\omega(t_0 - r)} \|\mathcal{U}(r)\|\, dr, \end{aligned}$$

and with (20),

$$\|T(t_0)\mathcal{U}_0 - S(t_0)\mathcal{U}_0\| \leq b(t_0)\|\mathcal{U}_0\|^2 \ \forall t_0 > 0 \quad \text{and} \quad \|\mathcal{U}_0\| < a_0.$$

From now on, we take a positive t_0 and assume, by changing b if necessary, that

$$b = b(t_0) \geq \frac{e^{\omega t_0} - 1}{8a_0 M^2}. \tag{21}$$

Then let

$$\varepsilon < \inf\left\{a_1, \frac{e^{\omega t_0}-1}{32M^3b}\right\}$$

and let $\alpha > 0$ be such that $\|\mathcal{U}_0\| < \alpha \Rightarrow \|\mathcal{U}(t)\| \leq \varepsilon$, for all positive t. Take $\lambda \in \sigma(S(t_0))$ with $|\lambda| = e^{\omega t_0}$. Note that $e^{\omega t_0}$ is the spectral radius of $S(t_0)$, i.e., $\lambda = e^{i\theta}e^{\omega t_0}$, with $\theta \in \mathbb{R}$.

Let N be an integer large enough so that

$$e^{\omega N t_0} \leq \frac{e^{\omega t_0 - 1}}{16M^3b\alpha} \tag{22}$$

and

$$|e^{iN\theta} - 1| \leq \frac{1}{12}. \tag{23}$$

Then, since $\lambda \in \sigma(S(t_0))$, we can find ξ and $\eta \in X$ such that $\|\xi\|^2 + \|\eta\|^2 \geq 1$ and $\|S(t_0)(\xi + i\eta) - \lambda(\xi + i\eta)\|_{X+iX}$ is arbitrarily small. To be more precise, we choose ξ and η in X such that $\|\eta\| \leq \|\xi\| = 1$ and satisfying

$$\|S(Nt_0)(\xi + i\eta) - \lambda^N(\xi + i\eta)\|_{X+iX} \leq \frac{1}{12}e^{\omega N t_0}.$$

Then we have

$$\|S(Nt_0)\xi - \mathcal{R}e(\lambda^N(\xi + i\eta))\| = \|S(Nt_0)\xi - e^{\omega N t_0}(\cos(N\theta)\xi - \sin(N\theta)\eta)\| \leq \frac{e^{\omega N t_0}}{12}.$$

Since it follows from (23) that

$$|\cos N\theta| \geq 1 - \frac{1}{12}$$

and

$$|\sin N\theta| \leq \frac{1}{12},$$

we obtain

$$\|S(Nt_0)\xi\| \geq \frac{3}{4}e^{\omega N t_0}.$$

Now, let

$$\delta = \frac{e^{\omega t_0} - 1}{16M^3be^{\omega N t_0}}.$$

It follows from (22) that $\delta \leq \alpha$, and thus if we take $\mathcal{U}_0 = \delta\xi$, we must have $\|T(t)\mathcal{U}_0\| \leq \varepsilon$ for all positive t. We will show that this is not satisfied, and more precisely that $\|\mathcal{U}(Nt_0)\| = \|T(Nt_0)\mathcal{U}_0\| \geq \varepsilon$.

For $0 \leq n \leq N$, we have

$$\begin{aligned}\mathcal{U}(nt_0) = T(nt_0)\mathcal{U}_0 &= S(nt_0)\mathcal{U}_0 + \sum_{k=0}^{n-1} S((n-1-k)t_0)[\mathcal{U}((k+1)t_0) - S(t_0)\mathcal{U}(kt_0)] \\ &= S(nt_0)\mathcal{U}_0 + \sum_{k=0}^{n-1} S((n-1-k)t_0)[T(t_0)\mathcal{U}(kt_0) - S(t_0)\mathcal{U}(kt_0)].\end{aligned}$$

Then, it is easy to show by induction, using (21) and (20), that

$$\|\mathcal{U}(nt_0) - S(nt_0)\mathcal{U}_0\| \leq \frac{\delta}{4} e^{\omega n t_0}, \quad 0 \leq n \leq N.$$

We deduce that

$$\|\mathcal{U}(Nt_0)\| \geq \|S(Nt_0)\mathcal{U}_0\| - \frac{\delta}{4} e^{\omega N t_0} \geq \frac{\delta}{2} e^{\omega N t_0} > \frac{e^{\omega t_0} - 1}{32 M^3 b} \geq \varepsilon,$$

and this proves the result. □

Acknowledgments. I am indebted to J. C. Saut who brought the subject to my attention, and J. Ginibre, who considerably simplified some of the proofs.

REFERENCES

[1] I. V. Barashenkov and V. G. Makhankov, *Soliton-like "bubbles" in a system of interacting bosons*, Phys. Lett. A, 128 (1988), pp. 52–56.

[2] ———, A. D. Gocheva, V. G. Makhankov, and I. V. Puzynin, *Stability of the soliton-like "bubbles"*, Physica D, 34 (1989), pp. 240–254.

[3] ——— and E. Y. Panova, *Stability and evolution of the quiescient and travelling solitonic bubbles*, Physica D, 69 (1993), pp. 114–134.

[4] ———, I. V. Puzynin, T. Zhanlav, and T. L. Boyadjiev, *Stability of the moving "bubbles" in a Bose condensate*, in Proc. Workshop on Solitons and Applications, V. G. Makhankov, V. K. Fedyanin, O. K. Pashaev, eds., World Scientific, Dubna, 1990, pp. 281–297.

[5] H. Berestycki and T. Cazenave, *Instabilité des états stationaires dans les équations de Schrödinger et de Klein-Gordon non linéaires*, C.R. Acad. Sci., Paris, Sér. I, Math., 293 (1981), pp. 489–492.

[6] ———, T. Gallouët, and O. Kavian, *Equations de champs scalaires euclidiens non linéaires dans le plan*, C. R. Acad. Sci., Paris, Sér. I, Math., 297 (1983), pp. 307–310.

[7] ——— and P. L. Lions, *Nonlinear scalar field equations* I: *existence of a ground state*, Arch. Rational Mech. Anal., 82 (1983), pp. 313–376.

[8] T. Cazenave and P. L. Lions, *Orbital stability of standing waves for some nonlinear Schrödinger equation*, Comm. Math. Phys., 85 (1982), pp. 549–561.

[9] A. de Bouard, *Etude de Quelques Propriétés d'Equations d'Ondes Non Linéaires Dispersives de Type Schrödinger*, Thèse, Orsay, France, 1992.

[10] L. E. Fraenkel, *Formulae for high derivatives of composite functions*, Math. Proc. Cambridge Philos. Soc., 83 (1978), pp. 159–165.

[11] M. Grillakis, *Linearized instability for nonlinear Schrödinger and Klein–Gordon equations*, Comm. Pure Appl. Math., 41 (1988), pp. 747–774.

[12] ———, *Analysis of the linearization around a critical point of an infinite dimensional Hamiltonian system*, Comm. Pure Appl. Math., 43 (1990), pp. 299–333.

[13] ———, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry* I, J. Funct. Anal., 74 (1987), pp. 160–197.

[14] D. B. Henry, J. F. Perez, and W. F. Wreszinski, *Stability theory of solitary waves solutions of scalar field equations*, Comm. Math. Phys., 85 (1982), pp. 351–361.

[15] J. Moser, *A rapidly convergent iteration method and non-linear partial differential equation*, Ann. Scuola Norm. Sup. Pisa, 3 (1966), pp. 265–315.

[16] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[17] M. Reed and B. Simon, *Method of Modern Mathematical Physics*, Vol. II, IV, Academic Press, New York, 1979.

[18] M. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.

[19] ———, *Lyapunov stability of ground states of nonlinear dispersive equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 583–600, May 1995

004

HEARING POINT MASSES IN A STRING*

ROBERT CARLSON†

Abstract. The spectral and inverse spectral theory for certain singular Sturm–Liouville problems is developed. These boundary value problems arise when considering the wave equation corresponding to a string with finitely many embedded point masses. These singular eigenvalue equations, their solutions, and the associated Hilbert space operators are constructed as limits of regular problems. The eigenfunctions of the singular problem are shown to be solutions of a regular eigenvalue problem with interior point conditions.

Expressions describing the distribution of large eigenvalues are developed. Algorithms are given for extracting information about the singularities from eigenvalues corresponding to one or two sets of boundary conditions. In the generic case a single spectrum determines the (unordered) set of lengths of the intervals separating the singularities.

Key words. Sturm–Liouville problem, inverse eigenvalue problem, inverse spectral theory

AMS subject classification. 34A55

1. Introduction.

The one-dimensional linear wave equation for the transverse vibrations of a string in the absence of exterior forces with constant tension 1 is [3]

$$\frac{\partial^2 u}{\partial t^2} = \frac{1}{\rho(x)} \frac{\partial^2 u}{\partial x^2},$$

subject to appropriate boundary conditions. "Hearing" the string density $\rho(x)$ is the problem of recovering ρ from the eigenvalues of the linear operator, which sends function f in its domain to f''/ρ. We are interested in direct and inverse spectral problems suggested by the physical model where the string density includes two contributions: a fixed, possibly inhomogeneous background density, and a finite family of masses with fixed "locations" and densities increasing to infinity. For technical convenience we use the reduction to Liouville normal form [5, p. 296] to remove the influence of the background density from the leading coefficient.

The main purpose of this work is to understand the effect of point masses on the sequence of eigenvalues of the string, and to extract information about the locations of the point masses from a single Sturm–Liouville problem. To make sense of point masses, we begin with ordinary differential operators of the form

$$L_\epsilon f = \frac{-1}{\rho_\epsilon(x)}[f''(x) + q(x)f(x)],$$

with ρ_ϵ integrable and real-valued, and q continuous and real-valued. The functions in the domain of the operator are assumed to satisfy the self-adjoint boundary conditions

$$af(0) - bf'(0) = 0, \tag{1.a}$$

$$cf(1) - df'(1) = 0,$$

*Received by the editors February 16, 1993; accepted for publication (in revised form) November 16, 1993.

†Department of Mathematics, University of Colorado, Colorado Springs, Colorado 80933.

where a, b, c, d are real. The limiting behavior is considered when the coefficients $\rho_\epsilon(x)$, $\epsilon > 0$, converge in the sense of distributions to

$$\rho_0(x) = 1 + \sum_{j=0}^{J-1} m_j \delta(x - x_j), \quad 0 < x_0 < \cdots < x_{J-1} < 1, \quad m_j > 0.$$

Under suitable hypotheses, the eigenvalues $\lambda_{\epsilon,k}$ converge to $\lambda_{0,k}$. The asymptotic behavior of $\lambda_{0,k}$ is described, and we give an algorithm for the recovery of the lengths $l_j = x_j - x_{j-1}$ from the tail $\{\lambda_{0,k}\}_{k>N}$ of a single eigenvalue sequence. This algorithm has a natural approximation when only a finite set of eigenvalues is known; it also provides an approach for approximating l_j when the masses have large but finite density.

The analysis of inverse spectral problems for second-order differential equations has led to a substantial body of literature. Early work on these problems was done by Borg [6], followed shortly thereafter by extensive work in the Soviet Union [14]–[18]. In particular, Krein [16] considered string equations with measures as coefficients; an extensive development of the spectral theory for such operators is in [15]. Inverse spectral problems are usually considered in the Liouville normal form

$$-y'' + q(x)y = \lambda y.$$

Unless $q(x)$ is even or otherwise constrained, one needs the spectra from two sets of boundary conditions, or one spectrum and a sequence of norming constants, to uniquely determine $q(x)$. Such inversion problems were solved for string equations with measures as coefficients by Krein [16].

More recently, new geometric ideas were developed for these problems by McKean and Trubowitz and their coworkers (see [21] and the references therein). Much of the recent work stressing irregular leading coefficients is motivated by geophysical models for oscillations of the earth. Hald [12] poses problems with an interior jump condition. Andersson [2] and Coleman and McLaughlin [11] consider problems of the form

$$(p^2 y')' + \lambda p^2 y = 0$$

with coefficients too irregular for the classic reduction to Liouville normal form. These and related results are surveyed in [19] and [8].

The investigation begins in §2, where convergence results show that the eigenvalue equation has meaningful solutions in the infinite density limit. In particular, Lemma 2.3 shows that the limiting solutions of the eigenvalue equation can be characterized as solutions of the "background" equation (no point masses) with interior point conditions at the point mass locations. These interior point conditions explicitly involve the mass at that point and the eigenvalue parameter.

Section 3 is devoted to an operator theoretic interpretation of the infinite density limit. For technical convenience we assume that 0 is not an eigenvalue of the "background" problem. We show that the limiting solutions of the eigenvalue equation from §2, which also satisfy the above boundary conditions, are the eigenfunctions of a self-adjoint Hilbert space operator with compact inverse. It is interesting that the Hilbert space on which the operator acts is not $L^2[0,1]$ but can be viewed as a finite-dimensional extension of this space [10].

Section 4 addresses the direct and inverse spectral theory for the limit operator. There is a fairly simple description of the eigenvalue asymptotics (Theorem 4.1), although they are more complex than in the usual case with regular coefficients. Turning

to the inverse spectral problem, we show that the eigenvalues are the zeros of an entire function $f(\lambda)$ of order $1/2$, which can thus be recovered, up to a constant factor, from the eigenvalues. An analysis of this entire function shows that in the case of Dirichlet boundary conditions, and generically for other separated boundary conditions, the unordered set of lengths $\{l_j = x_j - x_{j-1}\}$ is (constructively) determined by the tail $\{\lambda_k\}_{k=N}^{\infty}$ of an eigenvalue sequence from one set of boundary conditions. In the case of embedded point masses in an otherwise homogeneous string ($q(x) = 0$), additional results refine what was previously known about eigenvalue inversion for a string.

For the reader's convenience we mention some notational conventions. When convenient, $x(j)$ is written for x_j. Also, ∂_x is used for $\partial/\partial x$. A piecewise continuous function has limits $f(x_j^+) = \lim_{x \downarrow x_j} f(x)$ and $f(x_j^-) = \lim_{x \uparrow x_j} f(x)$.

2. Limiting solutions of the eigenvalue equation. Consider first the behavior, as $\epsilon \to 0$, $\epsilon > 0$, of the solutions of the initial value problem

$$L_\epsilon y_\epsilon = \lambda y_\epsilon, \quad y_\epsilon(0,\lambda) = b, \quad y_\epsilon'(0,\lambda) = a, \tag{2.a}$$

with

$$L_\epsilon y = \frac{-1}{\rho_\epsilon(x)}[\partial_x^2 y(x) + q(x)y(x)].$$

Write

$$\rho_\epsilon(x) = 1 + B_\epsilon(x) = 1 + \sum_j m_j b_{\epsilon,j}(x),$$

and assume that the functions $b_{\epsilon,j}(x) \geq 0$ are supported in the intervals $[x_j - \epsilon, x_j + \epsilon]$ and satisfy

$$\int_0^1 b_{\epsilon,j}(x)dx = 1.$$

The first lemma provides bounds for $|y_\epsilon(x,\lambda)|$ and $|y_\epsilon'(x,\lambda)|$ independent of ϵ. Let $M = \sum m_j$.

LEMMA 2.1. *If $y_\epsilon(x,\lambda)$ is the solution of* (2.a) *then, independent of ϵ, there is a function $K(a,b,\|q\|_\infty)$ such that*

$$|y_\epsilon(x,\lambda)| \leq K \ \exp(|Im\sqrt{\lambda}|x) \exp\left(\left|\sqrt{\lambda}\int_0^x B_\epsilon(t)dt\right|\right)$$

$$\leq K \ \exp(|Im\sqrt{\lambda}|x) \exp(|\sqrt{\lambda}M|)$$

and

$$|y_\epsilon'(x,\lambda)| \leq K[1 + |\sqrt{\lambda}|] \exp(|Im\sqrt{\lambda}|x) \exp(|\sqrt{\lambda}M|).$$

Proof. First introduce the solutions $C(x,t,\lambda)$, $S(x,t,\lambda)$ of the equation

$$-y''(x) + q(x)y = \lambda y(x),$$

which satisfy

$$C(t,t,\lambda) = 1, \quad C'(t,t,\lambda) = 0,$$

$$S(t,t,\lambda) = 0, \quad S'(t,t,\lambda) = 1.$$

We will write $C(x,\lambda)$ for $C(x,0,\lambda)$ and $S(x,\lambda)$ for $S(x,0,\lambda)$. The eigenvalue equation can be rewritten

$$-y_\epsilon'' + qy_\epsilon - \lambda y_\epsilon = B_\epsilon \lambda y_\epsilon.$$

This leads by variation of parameters to the equivalent integral equation

$$y_\epsilon(x,\lambda) = bC(x,\lambda) + aS(x,\lambda) - \lambda \int_0^x S(x,t,\lambda) B_\epsilon(t) y_\epsilon(t,\lambda)\, dt. \tag{2.b}$$

The function $S(x,t,\lambda)$ in turn can be represented as the unique solution of

$$S(x,t,\lambda) = \frac{\sin(\sqrt{\lambda}[x-t])}{\sqrt{\lambda}} - \int_t^x \frac{\sin(\sqrt{\lambda}[x-u])}{\sqrt{\lambda}} q(u) S(u,t,\lambda)\, du. \tag{2.c}$$

Defining the iterates

$$S_0(x,t,\lambda) = \frac{\sin(\sqrt{\lambda}[x-t])}{\sqrt{\lambda}},$$

$$S_{n+1}(x,t,\lambda) = \frac{\sin(\sqrt{\lambda}[x-t])}{\sqrt{\lambda}} + \int_t^x \frac{\sin(\sqrt{\lambda}[x-u])}{\sqrt{\lambda}} q(u) S_n(u,t,\lambda)\, du,$$

we find that

$$|S_0(x,t,\lambda)| \le \exp(|Im(\sqrt{\lambda}\,)||x-t|)$$

and, if $K_1 = \|q\|_\infty$,

$$|S_{n+1}(x,t,\lambda) - S_n(x,t,\lambda)| \le K_1 \int_t^x \exp(|Im(\sqrt{\lambda})||x-u|)|S_n(u,t,\lambda) - S_{n-1}(u,t,\lambda)|\, du.$$

By induction these inequalities yield the estimates

$$|S_{n+1}(x,t,\lambda) - S_n(x,t,\lambda)| \le \exp(|Im(\sqrt{\lambda})||x-t|)[K_1|x-t|]^n/n!,$$

and since $0 \le x,t \le 1$ there is a constant K_2 such that

$$|S(x,t,\lambda)| \le \exp(|Im(\sqrt{\lambda})||x-t|)\exp(K_1|x-t|]) \le K_2 \exp(|Im(\sqrt{\lambda})||x-t|).$$

Differentiation of (2.c) gives

$$\partial_t S(x,t,\lambda) = -\cos(\sqrt{\lambda}[x-t]) + \int_t^x \frac{\sin(\sqrt{\lambda}[x-u])}{\sqrt{\lambda}} q(u) \partial_t S(u,t,\lambda)\, du,$$

so that an argument like the previous one yields, with a new constant,

$$|\partial_t S(x,t,\lambda)| \le K_2 \exp(|Im(\sqrt{\lambda})||x-t|).$$

The estimates for $y_\epsilon(x,\lambda)$ and $y'_\epsilon(x,\lambda)$ are similarly obtained by using Picard iteration. Suppressing the ϵ for notational convenience, define

$$\phi_0(x,\lambda) = bC(x,\lambda) + aS(x,\lambda),$$

and for $n \ge 1$,

$$\phi_n(x,\lambda) = \phi_0(x,\lambda) - \lambda \int_0^x S(x,t,\lambda) B_\epsilon(t) \phi_{n-1}(t,\lambda)\, dt.$$

It is simple to see that there is a function $K_3(a,b)$ such that

$$|\phi_1 - \phi_0|(x,\lambda) \le K_3|\lambda| \exp(|Im\sqrt{\lambda}|x) \int_0^x B_\epsilon(t)dt$$

and then, by induction [20, p. 331],

$$|\phi_n - \phi_{n-1}|(x,\lambda) \le K_3|\lambda|^n \exp(|Im\sqrt{\lambda}|x) \left[\int_0^x B_\epsilon(t)dt\right]^n /n!.$$

Summing, we get the estimate for $|y_\epsilon(x,\lambda)|$. Differentiation of (2.b) and a parallel argument produces the estimates for $|y'_\epsilon(x,\lambda)|$. □

LEMMA 2.2. *As $\epsilon \to 0$, the functions $y_\epsilon(x,\lambda)$ converge uniformly on compact subsets of $[0,1] \times C$ to a continuous function $y_0(x,\lambda)$.*

Proof. Let $\eta \le \epsilon$. Using (2.b), we find

$$|y_\epsilon - y_\eta| \le \left|\lambda \int_0^x S(x,t,\lambda)B_\epsilon(t)[y_\epsilon(t,\lambda) - y_\eta(t,\lambda)]\ dt\right|$$

$$+ \left|\lambda \int_0^x S(x,t,\lambda)[B_\epsilon(t) - B_\eta(t)]y_\eta(t,\lambda)\ dt\right|.$$

The last integral is estimated first. Since

$$y_\eta(t,\lambda) = y_\eta(x_j - \epsilon,\lambda) + \int_{x_j-\epsilon}^t \partial_u y_\eta(u,\lambda)du,$$

we can write

$$\left|\lambda \int_0^x S(x,t,\lambda)[B_\epsilon(t) - B_\eta(t)]y_\eta(t,\lambda)\ dt\right|$$
$$\le |\lambda| \sum_j m_j \left|\int_{x_j-\epsilon}^x S(x,t,\lambda)[b_{\epsilon,j}(t) - b_{\eta,j}(t)]y_\eta(x_j - \epsilon,\lambda)\ dt\right|$$
$$+ |\lambda| \sum_j m_j \left|\int_{x_j-\epsilon}^x S(x,t,\lambda)[b_{\epsilon,j}(t) - b_{\eta,j}(t)] \int_{x_j-\epsilon}^t \partial_u y_\eta(u,\lambda)\ du\ dt\right|.$$

On compact subsets of $[0,1] \times [0,1] \times C$, the second sum directly above goes to 0 as $\epsilon \to 0$ because of the uniform bounds on $|\partial_u y_\eta(u,\lambda)|$ derived in Lemma 2.1, and because the support of $b_{\epsilon,j}$ and $b_{\eta,j}$ are in $[x_j - \epsilon, x_j + \epsilon]$. The first sum goes to 0 because of the continuity of $S(x,t,\lambda)$ and the uniform bounds on $|y_\eta(t,\lambda)|$.

Thus for any $\xi > 0$ and any compact subset of $[0,1] \times [0,1] \times C$ we can find ϵ such that

$$\left|\lambda \int_0^x S(x,t,\lambda)[B_\epsilon(t) - B_\eta(t)]y_\eta(t,\lambda)\ dt\right| < \xi.$$

For such ϵ, ξ, we have

$$|y_\epsilon - y_\eta| \le \xi + K_2|\lambda| \int_0^x \exp(|\sqrt{\lambda}|)|B_\epsilon(t)|y_\epsilon(t,\lambda) - y_\eta(t,\lambda)|\ dt.$$

By Gronwall's inequality [13, p. 24],

$$|y_\epsilon - y_\eta| \le \xi \exp\Big(|K_2\lambda| \int_0^x \exp(|\sqrt{\lambda}|)|B_\epsilon(t)|\, dt\Big)$$

and so we have convergence to a continuous limit uniformly in compact subsets of $[0,1] \times C$. □

LEMMA 2.3. *The functions $y_0(x,\lambda)$ satisfy the differential equation*

$$-y_0'' + q(x)y_0 = \lambda y_0, \quad x \ne x_j \tag{2.d}$$

and interior point conditions

$$-\lambda m_j y_0(x_j,\lambda) = y_0'(x_j^+,\lambda) - y_0'(x_j^-,\lambda).$$

Moreover, $y_\epsilon'(x,\lambda)$ converges uniformly to $y_0'(x,\lambda)$ on compact subsets of $\{[0,1] \setminus \{x_j\}\} \times C$.

Proof. To see that the equations and interior point conditions are satisfied by the limiting solutions $y_0(x,\lambda)$, first observe that the functions $y_\epsilon(x,\lambda)$ satisfy the equation

$$-y_\epsilon'' + q(x)y_\epsilon = \lambda y_\epsilon, \quad |x - x_j| > \epsilon,$$

so that the uniform limit satisfies the differential equation away from x_j. Differentiation of (2.b) gives

$$y_\epsilon'(x,\lambda) = bC'(x,\lambda) + aS'(x,\lambda) - \lambda \int_0^x \partial_x S(x,t,\lambda) B_\epsilon(t) y_\epsilon(t,\lambda)\, dt.$$

Thus $y_0'(x_j^-,\lambda)$ and $y_0'(x_j^+,\lambda)$ exist and satisfy

$$y_0'(x_j^-,\lambda) = bC'(x_j,\lambda) + aS'(x_j,\lambda) - \lambda \sum_{i<j} m_i \partial_x S(x_j,x_i,\lambda) y_\epsilon(x_i,\lambda),$$

$$y_0'(x_j^+,\lambda) = bC'(x_j,\lambda) + aS'(x_j,\lambda) - \lambda \sum_{i\le j} m_i \partial_x S(x_j,x_i,\lambda) y_\epsilon(x_i,\lambda).$$

Subtracting gives the desired interior point conditions. □

Suppose that $y_1(x,\lambda,\epsilon)$, $y_2(x,\lambda,\epsilon)$ are solutions of (2.a) for $\epsilon > 0$, satisfying the initial conditions

$$y_1(0,\lambda,\epsilon) = 1, \quad y_2(1,\lambda,\epsilon) = 0,$$
$$y_1'(0,\lambda,\epsilon) = 0, \quad y_2'(1,\lambda,\epsilon) = 1.$$

Then the variation of parameters formula gives us solutions of the inhomogeneous equation

$$[L_\epsilon - \lambda]f = g, \quad f(0) = 0 = f'(0)$$

in the form

$$f(x,\lambda,\epsilon) = \int_0^x G_\epsilon(x,t,\lambda) g(t)\, dt$$

with

$$G_\epsilon(x,t,\lambda) = -y_1(x,\lambda,\epsilon)y_2(t,\lambda,\epsilon) + y_1(t,\lambda,\epsilon)y_2(x,\lambda,\epsilon).$$

As a consequence of Lemma 2.2 we have the next lemma.

LEMMA 2.4. *As $\epsilon \to 0$, the kernels $G_\epsilon(x,t,\lambda)$ converge uniformly on compact subsets of $[0,1]\times[0,1]\times C$ to a continuous function $G_0(x,t,\lambda)$ with $G_0(x,x,\lambda)=0$.*

3. The limit operator. As is well known, there are self-adjoint Hilbert space operators associated to the operators L_ϵ with boundary conditions (1.a). It will be convenient to have a Hilbert space operator interpretation of

$$L_0 = \lim_{\epsilon\to 0} L_\epsilon.$$

This is most conveniently done when there is no nontrivial solution of

$$-y'' + q(x)y = 0$$

that satisfies the boundary conditions (1.a). In this situation we will establish a convergence result for the inverses of the operators L_ϵ.

For continuous functions $g(x)$, consider the boundary value problems

$$\frac{-1}{\rho_\epsilon}[\partial_x^2 f_\epsilon + q(x)f_\epsilon] = g, \tag{3.a}$$

$$af(0) - bf'(0) = 0,$$

$$cf(1) - df'(1) = 0.$$

When 0 is not an eigenvalue, the self-adjoint operator $-\partial_x^2 + q(x)$ with these boundary conditions has an inverse which is an integral operator K with continuous kernel $K(x,t)$ [9, p. 192]. Rewriting the differential equation as

$$-\partial_x^2 f_\epsilon + q(x)f_\epsilon = g + B_\epsilon g$$

and applying K results in the equation

$$f_\epsilon(x) = \int_0^1 K(x,t)g(t)dt + \int_0^1 K(x,t)B_\epsilon g.$$

Since $K(x,t)$ and $g(x)$ are continuous, this equation will have a limiting form, and we have

LEMMA 3.1. *For each $g \in C[0,1]$, $\lim_{\epsilon\to 0} f_\epsilon(x)$ exists, and the operator $R : C[0,1] \to C[0,1]$ given by*

$$Rg = \lim_{\epsilon\to 0} f_\epsilon(x) = \int_0^1 K(x,t)g(t)dt + \sum_j m_j K(x,x_j)g(x_j) \tag{3.b}$$

is continuous and linear.

R will not extend continuously to $L^2[0,1]$, so we consider a new inner product on $C[0,1]$ given by

$$\langle f,g\rangle = \int_0^1 f(x)\bar{g}(x)dx + \sum_j m_j f(x_j)\bar{g}(x_j).$$

Define $\mathcal{H}$ to be the completion of $C[0,1]$ with respect to this inner product, i.e., $\mathcal{H} = L^2([0,1],\mu)$, where μ is the Lebesgue measure plus $\sum_j m_j\delta(x-x_j)$.

THEOREM 3.2. *R extends to a compact, self-adjoint operator on $\mathcal{H}$ with trivial null space. Every function f in the range of R is continuously differentiable except at $\{x_j\}$, where $f'(x_j^+)$ and $f'(x_j^-)$ exist.*

Proof. It suffices to test R on the continuous functions in $\mathcal{H}$. If g is continuous with $\|g\|_{\mathcal{H}} \leq 1$ then $\|g\|_2 \leq 1$ and the closure of

$$\left\{ \int_0^1 K(x,t)g(t)dt \mid \|g\|_{\mathcal{H}} \leq 1 \right\}$$

is compact in $C[0,1]$ and consists of functions whose derivative is (absolutely) continuous on [0,1]. Since the functionals given by evaluation at x_j are continuous on $C[0,1]$, the set also has compact closure in $\mathcal{H}$. The remaining term $\sum_j m_j K(x,x_j)g(x_j)$ is a continuous finite rank operator, so the R extends to a compact operator on $\mathcal{H}$. The structure of the second term in (3.b) [9, p. 192] gives the jump discontinuities of f at x_j.

To prove that R is self-adjoint, we simply compute

$$\begin{aligned}
\langle Rg, h\rangle &= \int_0^1 \Big(\int_0^1 K(x,t)g(t)dt + \sum_j m_j K(x,x_j)g(x_j) \Big) \bar{h}(x)dx \\
&\quad + \sum_k m_k \Big(\int_0^1 K(x_k,t)g(t)dt + \sum_j m_j K(x_k,x_j)g(x_j) \Big) \bar{h}(x_k) \\
&= \int_0^1 \int_0^1 K(x,t)g(t)dt\ \bar{h}(x)dx + \int_0^1 \sum_j m_j K(x,x_j)g(x_j)\bar{h}(x)dx \\
&\quad + \sum_k m_k \int_0^1 K(x_k,t)g(t)dt\ \bar{h}(x_k) + \sum_k m_k \sum_j m_j K(x_k,x_j)g(x_j)\bar{h}(x_k) \\
&= \langle g, Rh\rangle.
\end{aligned}$$

We can finish the argument by showing that when K has a trivial null space, so does R. Suppose that $Rg = 0$ for some $g \in \mathcal{H}$. Then we would have a sequence $\{g_n\}$ of continuous functions converging in $\mathcal{H}$ so that $Rg_n \to 0$ in $\mathcal{H}$. The sequence $\{g_n\}$ has a limit $h \in L^2[0,1]$, which must satisfy

$$\int_0^1 K(x,t)h(t)\ dt = \sum a_j K(x,x_j).$$

Since the left-hand side is continuously differentiable, all $a_j = 0$, and then $h = 0$. As a consequence, $g_n \to 0$ in H. □

Let L_0 be the densely defined inverse of R. From (3.b) we see that if $f \in E$ where $E = \{f$ has two continuous derivatives, the boundary conditions (1.a) are satisfied, and f and f'' vanish at all $x_j\}$, then $L_0 f = -\partial_x^2 f + q(x)f$. Further, a sequence of continuous functions g_n, which peak to a value of one at x_j and vanish for $|x - x_j| > 1/n$, will converge in $\mathcal{H}$ to a vector $v = 1_{x(j)}$ with $Rv = m_j K(x,x_j)$, so that $F = \text{span}\ \{K(x,x_j)\}$ is in the domain of L_0. Further, L_0 acting on $E + F$ has dense range in $\mathcal{H}$.

THEOREM 3.3. *The eigenfunctions ψ_k of $L_0 = R^{-1}$ are the functions satisfying the differential equations*

$$-\psi_k''(x) + q(x)\psi_k(x) = \lambda_k \psi_k(x), \quad x \neq x_j$$

and the boundary and interior point conditions

$$a\psi_k(0) - b\psi_k'(0) = 0, \quad c\psi_k(1) - d\psi_k'(1) = 0,$$

$$-\lambda_k m_j \psi_k(x_j) = \psi_k'(x_j^+) - \psi_k'(x_j^-).$$

Proof. Since 0 is not an eigenvalue of R, all eigenfunctions ψ_k of R, with eigenvalues $\mu_k = 1/\lambda_k$, are in the range of R and must be continuous. Using (3.b), we consider

$$\int_0^1 K(x,t)\psi_k(t)dt + \sum_j m_j K(x,x_j)\psi_k(x_j) = \mu_k\psi_k(x).$$

If we compute the difference of the derivatives from above and below x_j we find [9, p. 192]

$$-m_j\psi_k(x_j) = \mu_k[\psi_k'(x_j^+) - \psi_k'(x_j^-)].$$

Of course, for $x \neq x_j$ these functions satisfy $\mu_k[-\psi_k'' + q\psi_k] = \psi_k$.

Conversely, suppose ψ_k satisfies the equation and boundary and interior point conditions. Then the function $h = \psi_k - \lambda_k \sum_j m_j K(x,x_j)\psi_k(x_j)$ satisfies the differential equation

$$-h''(x) + q(x)h(x) = \lambda_k\psi_k, \quad x \neq x_j,$$

the boundary conditions, and has a continuous derivative on [0,1]. Thus h satisfies the equation for all $x \in [0,1]$ and we have

$$\psi_k - \lambda_k \sum_j m_j K(x,x_j)\psi_k(x_j) = \lambda_k \int_0^1 K(x,t)\psi_k(t)\,dt,$$

which is what we wanted to show. □

To establish the convergence of the eigenvalues of L_ϵ to those of L_0, we first provide a lemma using the technique of [21, p. 30]. Suppose y_0 is as described in Lemma 2.3, satisfying the initial conditions

$$y(0) = b, \quad y'(0) = a. \tag{3.c}$$

LEMMA 3.4. *If λ is an eigenvalue for L_0, then*

$$\partial_\lambda[cy_0(1) - dy_0'(1)] \neq 0.$$

Proof. For $x \neq x_j$, differentiate the equation

$$-y_0'' + q(x)y_0 = \lambda y_0 \tag{3.d}$$

with respect to λ to get

$$-[\partial_\lambda y_0]'' + q(x)\partial_\lambda y_0 = y_0 + \lambda\partial_\lambda y_0. \tag{3.e}$$

Multiply (3.c) by $\partial_\lambda y_0$, equation (3.d) by y_0, and subtract to get

$$-[(\partial_\lambda y_0)'y_0 - (\partial_\lambda y_0)y_0']' = y_0^2.$$

Integration yields

$$
\begin{aligned}
&-[(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](x_0^-)\\
&-\sum_{j=1}^{J-1}\Big([(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](x_j^-)-[(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](x_{j-1}^+)\Big)\\
&-[(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](1)+[(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](x_{J-1}^+)=\int_0^1 y_0^2
\end{aligned}
$$

or

$$
\begin{aligned}
&-\sum_{j=0}^{J-1}\Big([(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](x_j^-)-[(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](x_j^+)\Big)\\
&-[(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](1)=\int_0^1 y_0^2.
\end{aligned}
$$

Differentiation of the interior point conditions

$$-\lambda m_j y_0(x_j)=y_0'(x_j^+)-y_0'(x_j^-)$$

gives

$$-m_j y_0(x_j)-\lambda m_j\partial_\lambda y_0(x_j)=\partial_\lambda y_0'(x_j^+)-\partial_\lambda y_0'(x_j^-).$$

Placing this into the above, we find

$$
\begin{aligned}
&\sum_{j=0}^{J-1}\Big([(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](x_j^-)-[(\partial_\lambda y_0)'y_0-(\partial_\lambda y_0)y_0'](x_j^+)\Big)\\
&=\sum_{j=0}^{J-1}\Big([(\partial_\lambda y_0)'(x_j^-)-(\partial_\lambda y_0)'(x_j^+)]y_0(x_j)-\partial_\lambda y_0(x_j)[y_0'(x_j^-)-y_0'(x_j^+)]\Big)\\
&=\sum_{j=0}^{J-1}\Big([m_j y_0(x_j)+\lambda m_j\partial_\lambda y_0(x_j)]y_0(x_j)-\partial_\lambda y_0(x_j)[\lambda m_j y_0(x_j)]\Big)\\
&=\sum_{j=0}^{J-1}\Big(m_j y_0^2(x_j)\Big).
\end{aligned}
$$

Thus we have

$$[(\partial_\lambda y_0)'y_0-\partial_\lambda y_0 y_0'](1)=\int_0^1 y_0^2+\sum_{j=0}^{J-1}\Big(m_j y_0^2(x_j)\Big).$$

At an eigenvalue λ_n we must have $cy_0(1,\lambda_n)-d_0y'(1,\lambda_n)=0$. At least one of c,d is not zero. Suppose it is c. Then we find

$$[(d\partial_\lambda y_0)'(1,\lambda_n)-c\partial_\lambda y_0(1,\lambda_n)]y_0'(1,\lambda_n)=c\left[\int_0^1 y_0^2+\sum_{j=0}^{J-1}\Big(m_j y_0^2(x_j,\lambda_n)\Big)\right]$$

and so

$$\partial_\lambda[cy_0(1,\lambda_n)-dy_0'(1,\lambda_n)]\neq 0.$$

The argument is the same if $d \neq 0$.

Finally, let $\psi_{\epsilon,k}$ denote the eigenfunction of L_ϵ corresponding to the kth eigenvalue $\lambda_{\epsilon,k}$ and satisfying the initial conditions

$$\psi_{\epsilon,k}(0) = b, \quad \psi'_{\epsilon,k}(0) = a.$$

THEOREM 3.5. *For all* $k = 1, 2, \ldots,$

$$\lambda_k = \lim_{\epsilon \to 0} \lambda_{\epsilon,k}$$

and

$$\psi_k(x, \lambda_k) = \lim_{\epsilon \to 0} \psi_{\epsilon,k}(x, \lambda_{\epsilon,k})$$

uniformly for $x \in [0, 1]$.

Proof. We begin the proof with some bounds on the eigenvalues of L_ϵ. The operator L_ϵ is self-adjoint with respect to the inner product $\langle f, g \rangle_\epsilon = \int_0^1 f(x)\bar{g}(x)\rho_\epsilon(x)\, dx$. If $\phi(x)$ is one of its eigenfunctions with eigenvalue $\lambda_{\epsilon,n}$, then

$$\lambda_{\epsilon,n} = \frac{\langle L_\epsilon \phi, \phi \rangle_\epsilon}{\langle \phi, \phi \rangle_\epsilon} = \frac{\int_0^1 [-\phi''(x) + q(x)\phi(x)]\bar{\phi}(x)\, dx}{\int_0^1 \rho_\epsilon(x)\phi(x)\bar{\phi}(x)\, dx}.$$

If the eigenvalues of $-\partial_x^2 + q(x)$ with the boundary conditions (1.a) are nonnegative, then so is the right-hand side. Otherwise let ζ_1 be the smallest eigenvalue of $-\partial_x^2 + q(x)$ with the boundary conditions (1.a). Since $\rho_\epsilon(x) \geq 1$, if $\lambda_{\epsilon,n} < 0$ we have

$$\lambda_{\epsilon,n} \geq \frac{\int_0^1 [-\phi''(x) + q(x)\phi]\bar{\phi}(x)\, dx}{\int_0^1 \phi(x)\bar{\phi}(x)\, dx} \geq \zeta_1.$$

Thus there is an a priori lower bound for the eigenvalues of all L_ϵ, $\epsilon > 0$.

Now suppose that λ_1 is the smallest eigenvalue of L_0. By virtue of the description of the eigenfunctions given in Theorem 3.3, the dimension of the eigenspace is one. Since $\partial_\lambda [c y_0(1, \lambda_0) - d y_0'(1, \lambda_0)] \neq 0$, the function $c y_0(1, \lambda_0) - d y_0'(1, \lambda_0)$ has values of opposite sign in any neighborhood of λ_0. By virtue of the convergence results in Lemmas 2.2 and 2.3, for small ϵ the operator L_ϵ has an eigenvalue $\lambda_{\epsilon,i}$, the ith largest independent of ϵ by continuity of the eigenvalues, with $\lim_{\epsilon \to 0} \lambda_{\epsilon,i} = \lambda_1$. Let ζ be the minimum of 0 and the smallest eigenvalue of $-\partial_x^2 + q(x)$. If $i \neq 1$, then the inequalities $\lambda_{\epsilon,i} > \lambda_{\epsilon,1} > \zeta$ imply the convergence of a subsequence of $\lambda_{\epsilon,1}$ to an eigenvalue of L_0 which is smaller than λ_1, an impossibility. Thus $i = 1$. The rest of the eigenvalues are handled in similar fashion. □

4. Eigenvalue asymptotics and inverse problems. In this section, using the characterization of the eigenfunctions of L_0 in Theorem 3.3, the behavior of the eigenvalues λ_k of L_0 is described as $\lambda_k \to \infty$. Procedures are given for recovering the point masses and their positions from one or two spectra.

Given a basis for the solutions of the eigenvalue equation (2.d) on the intervals between the points x_j, we can find transition matrices which describe the change of representation as each x_j is crossed. For notational simplicity define $\omega = \sqrt{\lambda}$, $l_j = x_j - x_{j-1}$, $x_{-1} = 0$, and $x_J = 1$. The eigenfunctions are first represented in the bases

$$C(x, x_{j-1}, \lambda), \quad S(x, x_{j-1}, \lambda), \quad x_{j-1} < x < x_j.$$

Writing the coefficients of a solution y of the eigenvalue equation with respect the these bases as a column vector and using the interior point conditions

$$y(x_j^+,\lambda) = y(x_j^-,\lambda), \quad y'(x_j^+,\lambda) = y'(x_j^-,\lambda) - \omega^2 m_j y(x_j),$$

it is routine to compute that the transition matrix at x_j has the form

$$T_j = \begin{pmatrix} C(x_j,x_{j-1},\lambda) & S(x_j,x_{j-1},\lambda) \\ C'(x_j,x_{j-1},\lambda) & S'(x_j,x_{j-1},\lambda) \end{pmatrix} - \lambda m_j \begin{pmatrix} 0 & 0 \\ C(x_j,x_{j-1},\lambda) & S(x_j,x_{j-1},\lambda) \end{pmatrix}.$$

Of course, the transition across all of the points x_j is then represented by $T = T_{J-1}\dots T_0$.

In looking for eigenvalues we can assume that $y(0,\lambda) = b$, $y'(0,\lambda) = a$, and then the condition that the boundary conditions at both ends are satisfied is

$$(c \quad -d) \begin{pmatrix} C(x_J,x_{J-1},\lambda) & S(x_J,x_{J-1},\lambda) \\ C'(x_J,x_{J-1},\lambda) & S'(x_J,x_{J-1},\lambda) \end{pmatrix} T(\lambda) \begin{pmatrix} b \\ a \end{pmatrix} = 0.$$

It is well known [21, p. 13] that the following estimates are valid for all $x \in [0,1]$, $\lambda \in C$, and for all coefficients q satisfying a uniform bound:

$$|C(x,x_j,\lambda) - \cos(\omega[x-x_j])| \le \frac{K}{|\omega|}\exp(|Im(\omega)|), \tag{4.a}$$

$$|S(x,x_j,\lambda) - \sin(\omega[x-x_j])/\omega| \le \frac{K}{|\lambda|}\exp(|Im(\omega)|),$$

$$|C'(x,x_j,\lambda) + \omega\sin(\omega[x-x_j])| \le K\ \exp(|Im(\omega)|),$$

$$|S'(x,x_j,\lambda) - \cos(\omega[x-x_j])| \le \frac{K}{|\omega|}\exp(|Im(\omega)|).$$

Based on these estimates, we see that if

$$G(\omega) = \begin{pmatrix} 1 & 0 \\ 0 & \omega \end{pmatrix},$$

then

$$T_j = G(\omega)\left[\begin{pmatrix} \cos(\omega l_j) & \sin(\omega l_j) \\ -\sin(\omega l_j) & \cos(\omega l_j) \end{pmatrix} + O\left(\frac{1}{\omega}\exp(|Im(\omega)|)\right)\right]G^{-1}(\omega),$$

$$-m_j\omega G(\omega)\left[\begin{pmatrix} 0 & 0 \\ \cos(\omega l_j) & \sin(\omega l_j) \end{pmatrix} + O\left(\frac{1}{\omega}\exp(|Im(\omega)|)\right)\right]G^{-1}(\omega).$$

Thus T has the form

$$T = G(\omega)\left[\begin{pmatrix} 0 & 0 \\ B_1 & B_2 \end{pmatrix} m_0\dots m_{J-1}(-\omega)^J + O(\omega^{J-1}\exp(J|Im(\omega)|))\right]G^{-1}(\omega) \tag{4.b}$$

with

$$B_1(\omega) = \cos(\omega l_0)\sin(\omega l_1)\dots\sin(\omega l_{J-1}),$$

$$B_2(\omega) = \sin(\omega l_0)\dots\sin(\omega l_{J-1}).$$

Let $Q_T = G^{-1}TG$.

The location of the eigenvalues of L_0 is approximately given by the zero set of an elementary entire function.

THEOREM 4.1. *There is a sequence* $\{\alpha_N | N = 0, 1, 2, \ldots\}$ *satisfying* $N^2 \leq \alpha_N < [N+1]^2$ *such that for* N *sufficiently large, the number of eigenvalues* λ_k *of* L_0 *with* $\lambda_k < \alpha_N$ *agrees with the number of roots* $r < \alpha_N$ *of the entire function* $h(\lambda)$, *where*

$$h(\lambda) = \lambda^J \cos(\omega l_0) \frac{\sin(\omega l_1)}{\omega} \cdots \frac{\sin(\omega l_{J-1})}{\omega} \cos(\omega l_J) \quad \text{if} \quad bd \neq 0,$$

$$h(\lambda) = \lambda^J \frac{\sin(\omega l_0)}{\omega} \cdots \frac{\sin(\omega l_{J-1})}{\omega} \cos(\omega l_J) \quad \text{if} \quad b = 0,\ d \neq 0,$$

$$h(\lambda) = \lambda^J \cos(\omega l_0) \frac{\sin(\omega l_1)}{\omega} \cdots \frac{\sin(\omega l_J)}{\omega} \quad \text{if} \quad d = 0,\ b \neq 0,$$

$$h(\lambda) = \lambda^J \frac{\sin(\omega l_0)}{\omega} \cdots \frac{\sin(\omega l_J)}{\omega} \quad \text{if} \quad b = 0,\ d = 0.$$

Moreover, if r_k *is the root of* $h(\lambda)$ *closest to* λ_k, *then*

$$|\sqrt{\lambda_k} - \sqrt{r_k}| = O(\lambda_k^{-1/(2J+2)}).$$

Proof. Details will be provided for the cases $bd \neq 0$. The other cases are similar.

The proof is similar to the discussion in [21, p. 27]. We begin with the observation that if

$$|z - n\pi| > \beta > 0, \quad n = 0, 1, 2, \ldots,$$

then there is a constant c_β such that

$$\exp(|Im(z)|) < c_\beta |\sin(z)|,$$

and similarly, if

$$|z - [n\pi + \pi/2]| > \beta > 0, \quad n = 0, 1, 2, \ldots,$$

then

$$\exp(|Im(z)|) < c_\beta |\cos(z)|.$$

Based on the explicit form of $h(\lambda)$ one sees easily that α_N can be chosen so that there is a $\beta > 0$ with $|\alpha_N - r| > \beta$ whenever $h(r) = 0$.

The condition that there is an eigenvalue at $\lambda = \omega^2$ is the equation

$$cy(1,\lambda) - dy'(1,\lambda) = (\, c \quad -d\,) G \begin{pmatrix} \cos(\omega l_J) & \sin(\omega l_J) \\ -\sin(\omega l_J) & \cos(\omega l_J) \end{pmatrix} Q_T G^{-1} \begin{pmatrix} b \\ a \end{pmatrix} = 0.$$

Using (4.b) we see that in case $db \neq 0$ this equation has the form

$$\text{(4.c)} \quad cy(1,\lambda) - dy'(1,\lambda) = -db\cos(\omega l_J) B_1(\omega) \omega^{J+1} + O(\omega^J \exp([J+1]|Im(\omega)|)) = 0.$$

On the contour $|\lambda| = \alpha_N$ there is a constant K so that

$$|cy(1,\lambda) - dy'(1,\lambda) + dbh(\lambda)| < \left| \frac{K}{\omega} h(\lambda) \right|.$$

For N sufficiently large, Rouche's theorem yields the first part of the theorem.

For more precise information about eigenvalue location, let Z_j denote the set of roots of $\sin(\omega l_j)$ or $\cos(\omega l_j)$, depending on which trigonometric function appears in $h(\lambda)$. Then

$$|\sin(\omega l_j)| \geq \frac{2}{\pi}\operatorname{dist}(\omega, Z_j),$$

and similarly for cosine, so that

$$h(\lambda) \geq K_1 \omega^{J+1} \prod_{j=0}^{J} \operatorname{dist}(\omega, Z_j).$$

Again, by making use of (4.c), we have, for large ω,

$$|cy(1,\lambda) - dy'(1,\lambda) + dbh(\lambda)| < |K\omega^J|.$$

Thus at any eigenvalue λ_k we must have

$$\prod_{j=0}^{J} \operatorname{dist}(\sqrt{\lambda_k}, Z_j) \leq \frac{K_2}{\sqrt{\lambda_k}}.$$

Letting $Z = \cup_{j=0}^{J} Z_j$ implies

$$\operatorname{dist}(\sqrt{\lambda_k}, Z) \leq K_2 \lambda_k{}^{-1/(2j+2)}. \qquad \square$$

Among the set $\mathcal{L}$ of $(J+1)$-*tuples* l_j such that $0 < l_0 < \cdots < l_J$ and $\sum_j l_j = 1$, it will be convenient to distinguish a certain subset. If the vectors $B_i = (\beta_{i,0}, \ldots, \beta_{i,J})$, $i = 1, 2$ are distinct and where $\beta_{i,j} = \pm 1$, then the subset of $\mathcal{L}$ satisfying an equation of the form

$$\sum_j \beta_{1,j} l_j = \sum_j \beta_{2,j} l_j$$

is a hypersurface, and the set $U \subset \mathcal{L}$ consisting of those $(J+1)-$ *tuples* that satisfy no equation of this form is an open dense set.

THEOREM 4.2. *If $b = d = 0$, or if $b^2 + d^2 \neq 0$ and the size ordered lengths l_j, $j = 0, \ldots, J$ are an element of U, then the lengths l_j are determined by the eigenvalues $\{\lambda_k : k > K\}$ of L_0.*

Proof. Again, we emphasize the case $bd \neq 0$, leaving the minor modification needed for the other cases to the reader. To begin, we assume that all eigenvalues λ_k of L_0 are known. From (4.c) (and its analog for the other boundary conditions), we see that the entire function $cy(1,\lambda) - dy'(1,\lambda)$, whose zeros are precisely the eigenvalues of L_0, has order $1/2$. Thus [1, p. 207]

$$cy(1,\lambda) - dy'(1,\lambda) = K_1 \prod (1 - \lambda/\lambda_k),$$

where K_1 is a nonzero constant (recall that zero is not an eigenvalue).

From the function $f(\lambda) = \prod(1 - \lambda/\lambda_k)$, we can recover $h(\lambda)$ up to a constant multiple. Rewriting $h(\lambda)$ with exponentials, we find that

$$h(\lambda) = -\left[\frac{\omega}{2i}\right]^{J+1} \sum_B \alpha_B \exp\left(i\omega \sum_j \beta_{j,B} l_j\right), \tag{4.d}$$

where the index set B consists of vectors $(\beta_0, \dots, \beta_J)$ with $\beta_j = \pm 1$, and $\alpha_B = \pm 1$.

For real ν, consider the integrals

$$\text{(4.e)} \qquad \lim_{X\to\infty} X^{-1} \int_1^X \exp(-i\nu\omega)\omega^{-\mu} f(\lambda)\, d\omega.$$

From (4.c) and (4.d) these limits will all be zero for $\mu > J+1$, and there will be a finite set of ν with nonzero limits for $\mu = J+1$ corresponding to the coefficients of $dbh(\lambda)$ in the representation (4.d). Thus the eigenvalues of L_0 determine J and they determine $h(\lambda)$ up to a constant multiple.

The assumption that the size ordered set of lengths is in U means that the terms $\nu_B = \sum_j \beta_{j,B} l_j$ appearing in (4.d) are all distinct. Consequently, the set of these ν_B will be exactly those ν for which a nonzero limit appears in (4.e).

The lengths l_j can be found as follows. The largest value is $\nu_B = \sum_j l_j$. The next largest is $\sum_j l_j - 2l_0$. Thus l_0 can be found. Proceeding inductively, after finding $l_0 < \cdots < l_K$ and discarding those values of ν_B of the form $\sum_j l_j - 2\sum_{k=0}^{K} \pm l_k$, the next largest is $\sum_j l_j - 2l_{k+1}$.

Note that the same procedure is applicable if the eigenvalues $\lambda_1, \dots, \lambda_K$ are unknown: Let $\mu_1, \dots, \mu_K$ be real and nonzero, and consider the function

$$g(\lambda) = \prod_{k=1}^{K}(1 - \lambda/\mu_k) \prod_{k>K}(1 - \lambda/\lambda_k) = \frac{\prod_{k=1}^{K}(1-\lambda/\mu_k)}{\prod_{k=1}^{K}(1-\lambda/\lambda_k)} f(\lambda).$$

For all values $\mu \geq J+1$, the integrals

$$\lim_{X\to\infty} X^{-1} \int_1^X \exp(-i\nu\omega)\omega^{-\mu} g(\lambda)\, d\omega, \quad \nu \text{ real}$$

will yield the same values, up to a nonzero constant factor, since

$$\frac{\prod_{k=1}^{K}(1-\lambda/\mu_k)}{\prod_{k=1}^{K}(1-\lambda/\lambda_k)} \to \prod_{k=1}^{K} \lambda_k/\mu_k$$

as $\lambda \to \infty$.

In case $b = 0 = d$, so that $h(\lambda)$ is a product of sines, the lengths can be recovered, without any restrictions, by using the location of the zeros of $h(\lambda)$. The location of the first positive zero gives the value of the largest l_j, and the multiplicity of the zero gives the number of intervals of that length. By successively examining the larger zeros of $h(\lambda)$ we can get all the lengths l_j. □

Without restrictions on the lengths l_j, some ambiguity can arise in the cases other than $b = 0 = d$. In particular, if $2l_1 + 2l_2 = 1$, the lengths $(2l_1, l_2, l_2)$ and $(l_1, l_1, 2l_2)$ will not be distinguished in case we have the boundary conditions $b = 0, d \neq 0$, which can be seen by observing that

$$\sin(2l_1\omega)\sin(l_2\omega)\cos(l_2\omega) = 2\sin(2l_1\omega)\sin(2l_2\omega) = \sin(l_1\omega)\sin(2l_2\omega)\cos(l_1\omega).$$

Since the proof of Theorem 4.2 only depended on the leading asymptotics of the function $cy(1,\lambda) - dy'(1,\lambda)$, the information about the lengths l_j and thus the points x_j was obtained without necessarily knowing $q(x)$. The remaining results address the case of point masses embedded in an otherwise homogeneous string, so that we can

assume $q(x) = 0$. It is convenient to express solutions of the eigenvalue equation in the basis $\cos(\omega x)$, $\sin(\omega x)/\omega$ on each interval (x_j, x_{j+1}). A straightforward computation shows that the transition matrix at x_j has the form

$$\mathcal{T}_j = G(\omega)\left[I + \frac{m_j\omega}{2}A_j(\omega)\right]G^{-1}(\omega),$$

where

$$A_j(\omega) = \begin{pmatrix} \sin(2\omega x_j) & 1-\cos(2\omega x_j) \\ -1-\cos(2\omega x_j) & -\sin(2\omega x_j) \end{pmatrix}.$$

Let $\mathcal{T} = \mathcal{T}_{J-1}\dots\mathcal{T}_0$ and let

$$Q_{\mathcal{T}}(\omega) = G^{-1}(\omega)\mathcal{T}(\omega)G(\omega) = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}.$$

Expanding the product for $\mathcal{T}$ gives

$$Q_{\mathcal{T}}(\omega) = I + \sum_j \frac{m_j\omega}{2}\begin{pmatrix} \sin(2\omega x_j) & 1-\cos(2\omega x_j) \\ -1-\cos(2\omega x_j) & -\sin(2\omega x_j) \end{pmatrix} + \cdots.$$

Now the condition that λ is an eigenvalue can be written as

(4.f)

$$\begin{aligned} f(\lambda) &= (c \quad -d)\,G(\omega)\begin{pmatrix} \cos(\omega) & \sin(\omega) \\ -\sin(\omega & \cos(\omega) \end{pmatrix}G^{-1}(\omega)G(\omega)Q_{\mathcal{T}}G^{-1}(\omega)\begin{pmatrix} b \\ a \end{pmatrix} \\ &= \frac{ac}{\omega}[\cos(\omega)Q_{12} + \sin(\omega)Q_{22}] + bc[\cos(\omega)Q_{11} + \sin(\omega)Q_{21}] \\ &\quad + ad[\sin(\omega)Q_{12} - \cos(\omega)Q_{22}] + bd\omega[\sin(\omega)Q_{11} - \cos(\omega)Q_{21}] = 0. \end{aligned}$$

Since $q(x) = 0$ is known and $\mathcal{T}(0) = I$, (4.f) gives

$$f(0) = (c \quad -d)\begin{pmatrix} \cos(\omega) & \sin(\omega)/\omega \\ -\omega\sin(\omega) & \cos(\omega) \end{pmatrix}\begin{pmatrix} b \\ a \end{pmatrix}.$$

Also,

$$f(\lambda) = K_1\prod(1-\lambda/\lambda_k),$$

so that $K_1 = f(0)$, and knowing the eigenvalues of L_0 determines $f(\lambda)$.

If we expand the product for $Q_{\mathcal{T}}$, gathering the terms with like powers of ω, we see that $f(\lambda)$ can be written as a polynomial in ω with coefficients which are generalized trigonometric polynomials with the form $\sum_k[\alpha_k\cos(f_k\omega) + \beta_k\sin(f_k\omega)]$. Analogous with the proof of Theorem 4.2, the frequencies and coefficients can be extracted from the sequence of eigenvalues by considering

$$\lim_{X\to\infty} X^{-1}\int_1^X \cos(\nu\omega)\omega^{-\mu}f(\lambda)\,d\omega, \quad \lim_{X\to\infty} X^{-1}\int_1^X \sin(\nu\omega)\omega^{-\mu}f(\lambda)\,d\omega.$$

Consider in particular the terms of lowest degree in ω with a nonconstant coefficient. These are

$$\frac{ac}{2}\sum_{j=0}^{J-1} m_j[\cos(\omega) - \cos(\omega[1-2x_j])], \quad ac \neq 0, \tag{4.g}$$

$$\frac{1}{2}\sum_{j=0}^{J-1} m_j[(ad-bc)\sin(\omega)-(bc+ad)\sin(\omega[1-2x_j])], \quad ac=0, \quad a^2+c^2\neq 0,$$

$$\frac{bd}{2}\sum_{j=0}^{J-1} m_j[\cos(\omega)+\cos(\omega[1-2x_j])], \quad a=c=0.$$

Since the distinct frequencies and the coefficients of the terms with distinct frequencies are determined from the eigenvalues, expressions (4.g) yield the following theorem.

THEOREM 4.3. *Suppose $q(x)=0$ and $ac\neq 0$ or $a=c=0$. Then the eigenvalues of L_0 determine the points x_j up to reflection about $x=1/2$, and they determine the sum of the masses at x_j and $1-x_j$. If $ac=0$ but not both a and c are zero, then the eigenvalues of L_0 determine the points x_j up to reflection about $x=1/2$ if the masses at x_j and $1-x_j$ are not equal. In this case the difference of the masses at x_j and $1-x_j$ is determined.*

Of course, by imposing additional constraints such as symmetry of the mass locations or restriction of the masses to $(0,1/2)$, this result will give conditions for the unique determination of the positions and masses from a single spectrum.

Next, while still restricting to $q(x)=0$, we consider the case when two spectra are given with different boundary conditions (that is, not differing only by scalings), with respective constants a_1,b_1,c_1,d_1 and a_2,b_2,c_2,d_2. We first consider the case when $a_1c_1a_2c_2\neq 0$. By a simple scaling we can assume $a_1=a_2$, $c_1=c_2$. Now notice that if, after this scaling, $b_1c_1+a_1d_1\neq b_2c_2+a_2d_2$, or equivalently $b_1/a_1+d_1/c_1\neq b_2/a_2+d_2/c_2$, then if we form the difference $f_1(\lambda)-f_2(\lambda)$ and use (4.f) with the expansion of Q_T in powers of ω, we can find $\sum_{j=0}^{J-1} m_j(bc+ad)\sin(\omega[1-2x_j])$. By knowing $m_{x(j)}+m_{1-x(j)}$ and $m_{x(j)}-m_{1-x(j)}$, we can find the masses at each x_j. Other cases are argued similarly. In summary, we have the following theorem.

THEOREM 4.4. *Suppose $q(x)=0$ and the eigenvalues of the operator L_0 are known for two linearly independent sets of boundary conditions. Then the masses m_j and locations x_j are determined if*

$$a_1c_1a_2c_2\neq 0, \quad b_1/a_1+d_1/c_1\neq b_2/a_2+d_2/c_2$$

or

$$a_1c_1=0, \quad b_1c_1+a_1d_1\neq 0, \quad a_2c_2\neq 0.$$

REFERENCES

[1] L. AHLFORS, *Complex Analysis,* McGraw-Hill, New York, 1966.

[2] L. ANDERSSON, *Inverse eigenvalue problems with discontinuous coefficients,* Inverse Problems, 4 (1983), pp. 353–397.

[3] S. ANTMAN, *The equations for large vibrations of strings,* Amer. Math. Monthly, May, 1980, pp. 359–370.

[4] V. BARCILON, *Inverse problems for vibrating elastic structures,* Proc. Eighth U.S. National Congress of Applied Mechanics, 1978, Western Periodicals, North Hollywood, CA, pp. 1–20.

[5] G. BIRKHOFF AND G.-C. ROTA, *Ordinary Differential Equations,* Blaisdell, Waltham, 1969.

[6] G. BORG, *Eine Umkehrung der Sturm–Liouville Eigenwertaufgabe,* Acta Math., 76 (1946), pp. 1–96.

[7] R. CARLSON, *An inverse spectral problem for Sturm–Liouville operators with discontinuous coefficients,* Proc. Amer. Math. Soc., 120 (1994), pp. 475–489.

[8] K. CHADAN AND P. SABATIER, *Inverse Problems in Quantum Scattering,* Springer-Verlag, New York, 1989.
[9] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations,* McGraw-Hill, New York, 1955.
[10] E. A. CODDINGTON, *Generalized resolutions of the identity for symmetric ordinary differential operators,* Ann. of Math., 68 (1958), pp. 378–392.
[11] C. F. COLEMAN AND J. R. MCLAUGHLIN, *Solution of the inverse spectral problem for an impedance with integrable derivative,* I *and* II, Comm. Pure Appl. Math., 46 (1993), pp. 145–212.
[12] O. HALD, *Discontinuous inverse eigenvalue problems,* Comm. Pure Appl. Math., 37 (1984), pp. 539–577.
[13] P. HARTMAN, *Ordinary Differential Equations,* John Wiley and Sons, Baltimore, MD, 1973.
[14] I. M. GELFAND AND B. M. LEVITAN, *On the determination of a differential equation by its spectral function,* Amer. Math. Soc. Transl. Ser. 2, 1 (1955), pp. 253–304.
[15] I. S. KAC AND M. G. KREIN, *On the spectral functions of the string,* Amer. Math. Soc. Transl. Ser. 2, 103 (1974), pp. 19–102.
[16] M. G. KREIN, *On inverse problems for an inhomogeneous string,* Dokl. Akad. Nauk. SSSR, 82 (1952), pp. 669–672.
[17] B. M. LEVITAN AND M. G. GASYMOV, *Determination of a differential equation by two of its spectra,* Russian Math. Surveys, 19 (1964), pp. 1–64.
[18] B. M. LEVITAN, *Inverse Sturm–Liouville Problems,* VNU Science Press, Utrecht, 1987.
[19] J. R. MCLAUGHLIN, *Analytical methods for recovering coefficients in differential equations from spectral data,* SIAM Rev., 28 (1986), pp. 53–72.
[20] R. NEWTON, *Scattering Theory of Waves and Particles,* McGraw-Hill, New York, 1966.
[21] J. PÖSCHEL AND E. TRUBOWITZ, *Inverse Spectral Theory,* Academic Press, Orlando, FL 1987.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 601–615, May 1995

005

EIGENVALUES OF THE FAR FIELD OPERATOR AND INVERSE SCATTERING THEORY*

DAVID COLTON† AND RAINER KRESS‡

Abstract. An eigenvalue-free region is determined for the far field operator corresponding to the scattering of time harmonic acoustic or electromagnetic waves by an inhomogeneous medium. In addition, a simple proof is given showing that for a nonabsorbing medium the far field operator is normal. These results are then used to derive a new method for solving the inverse scattering problem for acoustic and electromagnetic waves.

Key words. inverse scattering, far field operator

AMS subject classifications. 35P25, 35R30

1. Introduction.

The unitarity of the scattering matrix for nonabsorbing media is a basic result in scattering theory [1], [5]. For absorbing media, the scattering matrix is no longer a unitary operator and little is known about its spectrum. Closely related to the scattering matrix is the far field operator, which plays a central role in the dual space method for solving the inverse scattering problem [3]. Indeed, for nonabsorbing media, we will show in this paper that the unitarity of the scattering matrix follows from the proof of the normality of the far field operator. Since the far field operator is compact, zero is always an element of the spectrum, and of particular concern in the dual space method is the question of whether or not zero is an eigenvalue. In particular, if zero is an eigenvalue the dual space method does not work and this has led to various modifications of the far field operator to avoid this problem (c.f. [2], [4]).

In this paper, we shall consider an approach other than changing the far field operator for modifying the dual space method. In particular, instead of modifying the far field operator F, we shall consider $\lambda I - F$ where λ is not in the spectrum of F. Then, instead of considering modified far field operators, the corresponding dual space method makes use of the adjoint of the Herglotz operator discussed in §§5.5 and 7.4 of [3]. In this sense the present approach can be viewed as an extension of the method of Kirsch and Kress for solving the inverse obstacle problem (c.f. [3]) to the case of an inhomogeneous medium. We mention in passing that although in this paper we are only concerned with the scattering of acoustic and electromagnetic waves by an inhomogeneous medium, our ideas are also easily extendable to the case of obstacle scattering.

To accomplish the above program, we begin by establishing a relationship between the far field operator and Herglotz wave functions. As a consequence of this analysis, we obtain a region in the complex plane that is free of eigenvalues of the far field operator and, in the case of nonabsorbing media, a simple proof of the normality of the far field operator. We then combine these results with a denseness argument to derive a new method for solving the inverse scattering problem for acoustic and electromagnetic waves in an inhomogeneous medium. In our view, this new method

* Received by the editors May 22, 1993; accepted for publication October 11, 1993. This research was supported in part by a grant from the Air Force Office of Scientific Research.

† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19716.

‡ Institut für Numerische und Angewandte Mathematik, Universität Göttingen, 37083 Göttingen, Germany.

combines the attributes of both simplicity and flexibility and we hope to examine in the near future whether or not numerical experiments confirm this view.

2. The far field operator for acoustic waves in an inhomogeneous medium. We consider the scattering of a time harmonic acoustic plane wave by an inhomogeneous medium of compact support with complex valued refractive index $n \in C^1(\mathbb{R}^3), \operatorname{Im} n \geq 0$. Then, if $k > 0$ is the wave number, ω the frequency, and the incident field is given by

$$U^i(x,t) = \exp[i(kx \cdot d - \omega t)], \tag{2.1}$$

where $x \in \mathbb{R}^3$ and $d \in \mathbb{R}^3, |d| = 1$ is the direction of propagation, under appropriate assumptions [3] the amplitude u of the total field $U(x,t) = u(x)e^{-i\omega t}$ satisfies

$$\Delta u + k^2 n(x) u = 0 \quad \text{in } \mathbb{R}^3, \tag{2.2}$$

$$u(x) = e^{ikx\cdot d} + u^s(x), \quad x \in \mathbb{R}^3, \tag{2.3}$$

$$\lim_{r\to\infty} r\left(\frac{\partial u^s}{\partial r} - iku^s\right) = 0, \tag{2.4}$$

where $r = |x|$. From (2.2)–(2.4) it is easy to deduce [3] that $u^s(x) = u^s(x;d)$ has the asymptotic behavior

$$u^s(x;d) = \frac{\exp(ik|x|)}{|x|} u_\infty(\hat{x};d) + O\left(\frac{1}{|x|^2}\right) \tag{2.5}$$

as $|x| \to \infty$, where $\hat{x} = x/|x|$ and u_∞ is the *far field pattern* of the scattered field u^s.

Now let B denote an open ball (or some other domain with connected boundary) containing the support of $m := 1 - n$ and let ν be the exterior unit normal to ∂B. If Ω denotes the unit sphere in $\mathbb{R}^3$, then any solution of the Helmholtz equation

$$\Delta u + k^2 u = 0 \tag{2.6}$$

of the form

$$v(x) = \int_\Omega e^{ikx\cdot d} g(d)\, ds(d), \tag{2.7}$$

where $g \in L^2(\Omega)$ is called a *Herglotz wave function* with kernel g [3]. Solutions of (2.6) satisfying (2.4) are said to be radiating. Such solutions have the asymptotic behavior (2.5) [3].

LEMMA 2.1. *Let $v^s, w^s \in C^1(\mathbb{R}^3 \setminus B)$ be radiating solutions to the Helmholtz equation with far field patterns v_∞ and w_∞, respectively. Then*

$$\int_{\partial B} \left(v^s \frac{\partial \overline{w^s}}{\partial \nu} - \overline{w^s} \frac{\partial v^s}{\partial \nu}\right) ds = -2ik \int_\Omega v_\infty \overline{w_\infty}\, ds.$$

Proof. By Green's theorem, the value of the integral on the left-hand side remains the same if we replace ∂B by a sufficiently large sphere of radius R centered at the origin. The lemma now follows from

$$v^s(x) \frac{\partial \overline{w^s(x)}}{\partial |x|} - \overline{w^s(x)} \frac{\partial v^s(x)}{\partial |x|} = \frac{-2ik}{|x|^2} v_\infty(\hat{x}) \overline{w_\infty(\hat{x})} + O\left(\frac{1}{|x|^3}\right) \tag{2.8}$$

and letting R tend to infinity.

LEMMA 2.2. *Let $v^s \in C^1(\mathbb{R}^3\setminus B)$ be a radiating solution of the Helmholtz equation with far field pattern v_∞ and let w_h^i be a Herglotz wave function with kernel h. Then*

$$\int_{\partial B}\left(v^s\frac{\partial\overline{w_h^i}}{\partial\nu}-\overline{w_h^i}\frac{\partial v^s}{\partial\nu}\right)ds=4\pi\int_\Omega v_\infty\bar{h}\,ds\ .$$

Proof. From [3, p. 20], we have the integral representation

$$v_\infty(\hat{x})=\frac{1}{4\pi}\int_{\partial B}\left\{v^s(y)\frac{\partial}{\partial\nu}e^{-ik\hat{x}\cdot y}-\frac{\partial v^s}{\partial\nu}(y)e^{-ik\hat{x}\cdot y}\right\}ds(y)\ ,\quad \hat{x}\in\Omega\ . \tag{2.9}$$

Hence, since

$$w_h^i(x)=\int_\Omega e^{ik\,x\cdot d}\,h(d)\,ds(d)\ , \tag{2.10}$$

we have

$$\begin{aligned}&\int_{\partial B}\left(v^s(x)\,\frac{\partial\overline{w_h^i}}{\partial\nu}\,(x)-\overline{w_h^i(x)}\,\frac{\partial v^s}{\partial\nu}\,(x)\right)ds_{(x)} \\ &=\int_\Omega\overline{h(d)}\int_{\partial B}\left(v^s(x)\,\frac{\partial}{\partial\nu(x)}e^{-ik\,x\cdot d}-e^{-ik\,x\cdot d}\,\frac{\partial v^s(x)}{\partial\nu}\right)ds(x)\,ds(d) \\ &=4\pi\int_\Omega\overline{h(d)}v_\infty(d)\,ds(d).\end{aligned} \tag{2.11}$$

We now define the *far field operator* $F: L^2(\Omega)\to L^2(\Omega)$ corresponding to the far field pattern u_∞ by

$$(Fg)(\hat{x}):=\int_\Omega u_\infty(\hat{x};d)g(d)\,ds(d) \tag{2.12}$$

and note that F is a compact operator on $L^2(\Omega)$. The connection between the far field operator and Herglotz wave functions is given by the following theorem.

THEOREM 2.3. *Let v_g^i and v_h^i be Herglotz wave functions with kernels $g,h\in L^2(\Omega)$, respectively and let v_g, v_h be the solutions of* (2.2)–(2.4) *with $e^{ikx\cdot d}$ replaced by v_g^i and v_h^i, respectively. Then*

$$ik^2\int\int_B \operatorname{Im} n\,v_g\overline{v_h}\,dx=2\pi(Fg,h)-2\pi(g,Fh)-ik(Fg,Fh),$$

where $(\cdot,\cdot)$ denotes the inner product on $L^2(\Omega)$.

Proof. Let v_g^s, v_h^s denote the scattered fields corresponding to v_g and v_h, respectively and let $v_{g,\infty}, v_{h,\infty}$ be the corresponding far field patterns. Then, using Green's theorem, Lemma 2.1, and Lemma 2.2, we have

$$\begin{aligned}2ik^2\int\int_B \operatorname{Im} n\,v_g\overline{v_h}\,dx&=\int_{\partial B}\left(v_g\,\frac{\partial\overline{v_h}}{\partial\nu}-\overline{v_h}\,\frac{\partial v_g}{\partial\nu}\right)ds \\ &=\int_{\partial B}\left(v_g^s\,\frac{\partial\overline{v_h^s}}{\partial\nu}-\overline{v_h^s}\,\frac{\partial v_g^s}{\partial\nu}\right)ds+\int_{\partial B}\left(v_g^s\,\frac{\partial\overline{v_h^i}}{\partial\nu}-\overline{v_h^i}\,\frac{\partial v_g^s}{\partial\nu}\right)ds\end{aligned} \tag{2.13}$$

$$+\int_{\partial B}\left(v_g^i\frac{\partial\overline{v_h^s}}{\partial\nu}-\overline{v_h^s}\frac{\partial v_g^i}{\partial\nu}\right)ds$$
$$=-2ik\int_\Omega v_{g,\infty}\overline{v_{h,\infty}}\,ds+4\pi\int_\Omega v_{g,\infty}\bar{h}\,ds-4\pi\int_\Omega g\overline{v_{h,\infty}}\,ds$$
$$=-2ik(Fg,Fh)+4\pi(Fg,h)-4\pi(g,Fh)\,,$$

noting that Fg is the far field pattern corresponding to the incident field v_g^i.

COROLLARY 2.4. *Assume that* $\operatorname{Im} n\geq 0$. *Then, except for possibly zero,* F *has no real eigenvalues.*

Proof. Let $Fg=\lambda g$ with $\lambda\in\mathbb{R}$ and $\lambda\neq 0$. Then, choosing $h=g$ in Theorem 2.3, we have

$$ik^2\int\int_B \operatorname{Im} n\,|v_g|^2dx=4\pi i\operatorname{Im}(Fg,g)-ik(Fg,Fg)$$
$$=4\pi i\operatorname{Im}\lambda\|g\|^2-ik\|Fg\|^2. \tag{2.14}$$

Since $\operatorname{Im}\lambda=0$ we now have that $Fg=0$ and hence, since $\lambda\neq 0, g=0$.

COROLLARY 2.5. *Assume that* $\operatorname{Im} n=0$. *Then* F *is normal and hence has a countable number of eigenvalues.*

Proof. From Theorem 2.3 we have that

$$ik(Fg,Fh)=2\pi\{(Fg,h)-(g,Fh)\} \tag{2.15}$$

and hence

$$(g,ikF^*Fh)=2\pi\{(g,Fh)-(g,F^*h)\} \tag{2.16}$$

for all $g,h\in L^2(\Omega)$. We can now conclude that

$$ikF^*F=2\pi\{F-F^*\}. \tag{2.17}$$

By reciprocity [3, p. 53], we have

$$(F^*g)(\hat{x})=\int_\Omega\overline{u_\infty(d;\hat{x})}g(d)\,ds(d)=\int_\Omega\overline{u_\infty(-\hat{x};-d)}g(d)\,ds(d), \tag{2.18}$$

and hence if we define $R:L^2(\Omega)\to L^2(\Omega)$ by $(Rg)(d):=g(-d)$ we have that

$$F^*g=\overline{RFR\bar{g}}. \tag{2.19}$$

From this, observing that $(Rg,Rh)=(g,h)=(\bar{h},\bar{g})$ for all $g,h\in L^2(\Omega)$, we find that

$$(F^*g,F^*h)=(RFR\bar{h},RFR\bar{g})=(FR\bar{h},FR\bar{g}) \tag{2.20}$$

and hence, from (2.15),

$$ik(F^*g,F^*h)=2\pi\{(FR\bar{h},R\bar{g})-(R\bar{h},FR\bar{g})\}=2\pi\{(g,F^*h)-(F^*g,h)\}. \tag{2.21}$$

If we now proceed as in the derivation of (2.17), we find that

$$ikFF^*=2\pi\{F-F^*\} \tag{2.22}$$

and the proof is finished.

We note that (c.f. [1], [5]) the far field operator F is related to the scattering matrix S by

$$S = I + \frac{ik}{2\pi} F. \tag{2.23}$$

From (2.17) and (2.22) we now see that $SS^* = S^*S = I$, i.e., our analysis provides a rather simple proof of the well-known fact that the scattering matrix is a unitary operator.

3. The inverse scattering problem for acoustic waves in an inhomogeneous medium. The analysis of the preceding section suggests a new method for solving the inverse scattering problem for acoustic waves in an inhomogeneous medium. To see this, let Γ be a smooth surface contained in the interior of the support of m such that k^2 is not a Dirichlet eigenvalue for $-\Delta$ in the interior of Γ. For $\varphi \in L^2(\Gamma)$, define the single-layer potential

$$u_\varphi(x) := \int_\Gamma \Phi(x,y)\varphi(y)\, ds(y), \quad x \in \mathbb{R}^3 \setminus \Gamma, \tag{3.1}$$

where

$$\Phi(x,y) := \frac{1}{4\pi} \frac{e^{ik|x-y|}}{|x-y|}, \quad x \neq y, \tag{3.2}$$

and note that the far field pattern of u_φ is given by

$$u_{\varphi,\infty}(\hat{x}) = \frac{1}{4\pi} \int_\Gamma e^{-ik\,\hat{x}\cdot y} \varphi(y)\, ds(y). \tag{3.3}$$

Finally, for $\lambda \neq 0$ define $g \in L^2(\Omega)$ by

$$\lambda g(\hat{x}) := u_{\varphi,\infty}(\hat{x}) \tag{3.4}$$

and the Herglotz wave function v_g by

$$v_g(x) := \int_\Omega e^{ik\,x\cdot d} g(d)\, ds(d). \tag{3.5}$$

Now let u_∞ be the far field pattern corresponding to (2.2)–(2.4) and let $\sigma(F)$ denote the spectrum of the far field operator F corresponding to u_∞. For $\lambda \notin \sigma(F)$, let $g_\ell \in L^2(\Omega)$ be the unique solution of

$$Fg - \lambda g = \frac{1}{ki^{\ell+1}} Y_\ell \tag{3.6}$$

where Y_ℓ is a spherical harmonic order of $\ell, \ell = 0, 1, 2, \ldots$. Then, since the scattered field is uniquely determined by its far field pattern [3, p. 32], we have from (3.6) and unique continuation that if there exists a unique solution φ_ℓ of the integral equation $\lambda g_\ell = u_{\varphi,\infty}$ (c.f. [3, p. 128]) then

$$\int_\Omega u(x;d) g_\ell(d)\, ds(d) - v_{g_\ell}(x) - u_{\varphi_\ell}(x) = h_\ell^{(1)}(k|x|) Y_\ell(\hat{x}) \tag{3.7}$$

for $x \in \mathbb{R}^3 \setminus B$, where $h_\ell^{(1)}$ denotes the spherical Hankel function of the first kind of order $\ell, -L \leq \ell \leq L$. To solve the inverse scattering problem of determining n

from u_∞, one now tries to construct n from the overdetermined system (3.6), (3.7). If this system is formulated as an optimization problem as in [3, pp. 266–268], then the minimum value of the cost functional will be zero provided the following theorem is true, where

$$Y := \{w \in C^2(\bar{B}) : \Delta w + k^2 n w = 0 \text{ in } B\},$$

$$W := \left\{ \left(w - v_g - u_\varphi, \frac{\partial}{\partial \nu}(w - v_g - u_\varphi) \right) : \varphi \in L^2(\Gamma)\ ,\ \lambda g = u_{\varphi,\infty}\ ,\ w \in Y \right\},$$

and B is as defined in the previous section.

THEOREM 3.1. *Suppose $\lambda \notin \sigma(F)$. Then W is dense in $L^2(\partial B) \times L^2(\partial B)$.*

Proof. We first note that if j_0 is the spherical Bessel function of order zero then, since [3, p. 31]

$$\int_\Omega e^{ik\, d\cdot(x-y)} ds(d) = 4\pi j_0(k|x-y|), \tag{3.8}$$

we have that

$$v_g(x) = \frac{1}{\lambda} \int_\Gamma j_0(k|x-y|)\varphi(y)\, ds(y)\ . \tag{3.9}$$

Now suppose $a, b \in L^2(\Omega)$ satisfy

$$\int_{\partial B} \left\{ (w - v_g - u_\varphi)\bar{a} + \frac{\partial}{\partial \nu}(w - v_g - u_\varphi)\bar{b} \right\} ds = 0 \tag{3.10}$$

for all $w \in Y$ and $\varphi \in L^2(\Gamma)$. Then, setting $w = 0$ and defining

$$v(x) := \int_{\partial B} \left\{ \Psi(x,y)\overline{a(y)} + \frac{\partial}{\partial \nu(y)} \Psi(x,y)\overline{b(y)} \right\} ds(y)\ ,\quad x \in \mathbb{R}^3 \setminus \partial B, \tag{3.11}$$

where

$$\Psi(x,y) := \frac{1}{\lambda}\, j_0(k|x-y|) + \Phi(x,y), \tag{3.12}$$

we have from (3.10) that

$$\int_\Gamma \varphi v\, ds = 0 \tag{3.13}$$

for all $\varphi \in L^2(\Gamma)$. Hence $v = 0$ on Γ and, since k^2 is not a Dirichlet eigenvalue for $-\Delta$ in the interior of Γ, we have that $v = 0$ in the interior of Γ. By unique continuation, $v = 0$ in B and hence the L^2–jump relations for single- and double-layer potentials [3, p. 44] imply that

$$v_+ = \bar{b}, \quad \frac{\partial v}{\partial \nu}|_+ = -\bar{a} \quad \text{on } \partial B, \tag{3.14}$$

where $+$ denotes the limit taken from the exterior of B. Setting $\varphi = 0$ in (3.10) now implies that

$$\int_{\partial B} \left(\frac{\partial v}{\partial \nu}|_+\, w - v_+ \frac{\partial w}{\partial \nu} \right) ds = 0 \tag{3.15}$$

for all $w \in Y$.

We now define

$$v^s(x) := \int_{\partial B} \left\{ \Phi(x,y)\overline{a(y)} + \frac{\partial}{\partial \nu(y)} \Phi(x,y)\overline{b(y)} \right\} ds(y) \tag{3.16}$$

for $x \in \mathbb{R}^3 \setminus \bar{B}$ and

$$v^0(x) := \frac{1}{\lambda} \int_{\partial B} \left\{ j_0(k|x-y|)\overline{a(y)} + \frac{\partial}{\partial \nu(y)} j_0(k|x-y|)\overline{b(y)} \right\} ds(y) \tag{3.17}$$

for $x \in \mathbb{R}^3$ and note that from (3.8), v^0 is a Herglotz wave function

$$v^0(x) = \int_\Omega e^{ik\,x\cdot d}\, \tilde{g}(d)\, ds(d) \tag{3.18}$$

with kernel $\tilde{g} \in L^2(\Omega)$ defined by

$$\tilde{g}(d) := \frac{1}{4\pi\lambda} \int_{\partial B} \left\{ e^{-ik\,y\cdot d}\, \overline{a(y)} + \frac{\partial}{\partial \nu(y)} e^{-ik\,y\cdot d}\, \overline{b(y)} \right\} ds(y). \tag{3.19}$$

From (3.16) and (3.19) we now see that v^s has the far field pattern $v_\infty = \lambda \tilde{g}$.

In (3.15) we now set

$$w(x) = u^s(x; -\hat{z}) + e^{-ik\,x\cdot\hat{z}}, \tag{3.20}$$

where $\hat{z} \in \Omega$, and use Green's theorem to deduce that

$$\begin{aligned} &\int_{\partial B} \left(\frac{\partial v^s}{\partial \nu}(x)\, e^{-ik\,x\cdot\hat{z}} - v^s(x)\, \frac{\partial}{\partial \nu} e^{-ik\,x\cdot\hat{z}} \right) ds(x) \\ &+ \int_{\partial B} \left(\frac{\partial v^0}{\partial \nu}(x) u^s(x; -\hat{z}) - v^0(x)\, \frac{\partial u^s}{\partial \nu}(x; -\hat{z}) \right) ds(x) = 0. \end{aligned} \tag{3.21}$$

Substituting (3.18) into (3.21) and interchanging orders of integration, we now have from (2.9) and reciprocity that

$$\begin{aligned} v_\infty(\hat{z}) &= \int_\Omega \tilde{u}_\infty(-d; -\hat{z})\tilde{g}(d)\, ds(d) \\ &= \int_\Omega \tilde{u}_\infty(\hat{z}; d)\tilde{g}(d)\, ds(d), \end{aligned} \tag{3.22}$$

i.e.,

$$F\tilde{g} = v_\infty = \lambda \tilde{g}. \tag{3.23}$$

Since $\lambda \notin \sigma(F)$ we have that $\tilde{g} = 0$. This implies $v_\infty = 0$ and hence $v^s = 0$ in $\mathbb{R}^3 \setminus \bar{B}$ and $v^0 = 0$ in $\mathbb{R}^3$. Hence, $v = 0$ in $\mathbb{R}^3 \setminus \bar{B}$ and from (3.14) we now have that $a = b = 0$. The theorem is now proved.

4. The far field operator for electromagnetic waves in an inhomogeneous medium. We now consider the scattering of a time harmonic electromagnetic plane wave by an inhomogeneous medium of compact support. In particular, factoring out the time harmonic term $e^{-i\omega t}$ and letting $n \in C^1(\mathbb{R}^3)$, $\operatorname{Im} n \geq 0$, denote the

refractive index, we have that the (normalized) electric field E and magnetic field H satisfy [3, p. 239] the Maxwell equations

$$\operatorname{curl} E - ikH = 0, \quad \operatorname{curl} H + iknE = 0 \quad \text{in } \mathbb{R}^3, \tag{4.1}$$

where $k > 0$ is the wave number and $E = E^i + E^s, H = H^i + H^s$ with (E^i, H^i) being the incident electromagnetic field and (E^s, H^s) the scattered electromagnetic field. More specifically, (E^s, H^s) satisfies the Silver–Müller radiation condition

$$\lim_{r\to\infty} (H^s \times x - rE^s) = 0 \tag{4.2}$$

and we will assume that the incident field (E^i, H^i) is given by

$$\begin{aligned} E^i(x) &= E^i(x; d, p) = \frac{i}{k} \operatorname{curl}\operatorname{curl} p\, e^{ik\,x\cdot d} = ik\,(d \times p) \times d\, e^{ik\,x\cdot d}, \\ H^i(x) &= H^i(x; d, p) = \operatorname{curl} p\, e^{ik\,x\cdot d} = ik\, d \times p\, e^{ik\,x\cdot d}, \end{aligned} \tag{4.3}$$

where d is the direction, $d \in \mathbb{R}^3, |d| = 1$, and p is the polarization, $p \in \mathbb{R}^3$, of the incident plane wave.

From (4.1) and (4.2) it is easy to deduce [3] that the scattered electric field $E^s(x) = E^s(x; d, p)$ has the asymptotic behavior

$$E^s(x; d, p) = \frac{\exp(ik|x|)}{|x|} \left\{ E_\infty(\hat{x}; d, p) + 0\left(\frac{1}{|x|^2}\right) \right\} \tag{4.4}$$

as $|x| \to \infty$, where E_∞ is the *electric far field pattern*. We again let $m := 1 - n$, assume that m has compact support, and let B denote an open ball (or some other domain with connected boundary) containing the support of m with ν the exterior unit normal to ∂B. A solution to the Maxwell equations

$$\operatorname{curl} E - ikH = 0, \quad \operatorname{curl} H + ikE = 0 \tag{4.5}$$

of the form

$$\begin{aligned} E(x) &= \int_\Omega E^i(x; d, g(d))\, ds(d), \\ H(x) &= \int_\Omega H^i(x; d, g(d))\, ds(d), \end{aligned} \tag{4.6}$$

where $g \in T^2(\Omega) := \{g : \Omega \to \mathbb{C}^3 : g \in L^2(\Omega), g \cdot \nu = 0\}$ with ν being the exterior unit normal to the unit sphere Ω, is called an *electromagnetic Herglotz pair* with kernel g [3]. Solutions of (4.5) satisfying (4.2) are said to be radiating. Such solutions are easily seen to have the asymptotic behavior (4.4) [3].

LEMMA 4.1. *Let $E_1^s, H_1^s \in C^1(\mathbb{R}^3 \setminus \bar{B})$ and $E_2^s, H_2^s \in C^1(\mathbb{R}^3 \setminus \bar{B})$ be radiating solutions of the Maxwell equations with electric far field patterns $E_{1,\infty}, E_{2,\infty}$, respectively. Then*

$$\int_{\partial B} (\nu \times E_1^s \cdot \operatorname{curl} \overline{E_2^s} - \nu \times \overline{E_2^s} \cdot \operatorname{curl} E_1^s)\, ds = -2ik \int_\Omega E_{1,\infty} \cdot \overline{E_{2,\infty}}\, ds \ .$$

Proof. By the vector Green's theorem, the value of the integral on the left-hand side remains the same if we replace ∂B by a sufficiently large sphere of radius R centered at the origin. The lemma now follows from

$$\begin{aligned} &\nu(x) \times E_1^s(x) \cdot \operatorname{curl} \overline{E_2^s(x)} - \nu(x) \times \overline{E_2^s(x)} \cdot \operatorname{curl} E_1^s(x) \\ &= \frac{-2ik}{|x|^2} E_{1,\infty}(\hat{x}) \overline{E_{2,\infty}(\hat{x})} + O\left(\frac{1}{|x|^3}\right) \end{aligned} \tag{4.7}$$

and letting R tend to infinity.

LEMMA 4.2. *Let $E^s, H^s \in C^1(\mathbb{R}^3 \setminus B)$ be a radiating solution to the Maxwell equations with electric far field pattern E_∞ and let E_h^i, H_h^i be an electromagnetic Herglotz pair with kernel h. Then*

$$\int_{\partial B} (\nu \times E^s \cdot \operatorname{curl} \overline{E_h^i} - \nu \times \overline{E_h^i} \cdot \operatorname{curl} E^s)\, ds = -4\pi i k \int_\Omega E_\infty \cdot \overline{h}\, ds.$$

Proof. From [3, p. 157], we have the integral representation

$$E_\infty(\hat{x}) = \frac{ik}{4\pi} \hat{x} \times \int_{\partial B} \{\nu(y) \times E(y) + (\nu(y) \times H(y)) \times \hat{x}\} e^{-ik\hat{x}\cdot y}\, ds(y). \tag{4.8}$$

Furthermore, since

$$E_h^i(x) = \int_\Omega E^i(x; d, h(d))\, ds(d)\,, \tag{4.9}$$

we have

$$\begin{aligned} &\int_{\partial B} \nu(x) \times E^s(x) \cdot \operatorname{curl} \overline{E_h^i(x)}\, ds(x) \\ &= -k^2 \int_{\partial B} \nu(x) \times E^s(x) \cdot \int_\Omega d \times \overline{h(d)}\, e^{-ik\,x\cdot d} ds(d)\, ds(x) \\ &= k^2 \int_\Omega \overline{h(d)} \cdot \left(d \times \int_{\partial B} \nu(x) \times E^s(x)\, e^{-ik\,x\cdot d} ds(x) \right) ds(d) \end{aligned} \tag{4.10}$$

and

$$\begin{aligned} &-\int_{\partial B} \nu(x) \times \overline{E_h^i(x)} \cdot \operatorname{curl} E^s(x)\, ds(x) = ik \int_{\partial B} \overline{E_h^i(x)} \cdot \nu(x) \times H^s(x)\, ds(x) \\ &= k^2 \int_{\partial B} \left(\int_\Omega (d \times \overline{h(d)}) \times d\, e^{-ik\,x\cdot d}\, ds(d) \right) \cdot (\nu(x) \times H^s(x))\, ds(x) \\ &= k^2 \int_\Omega \overline{h(d)} \cdot \left(d \times \int_{\partial B} (\nu(x) \times H^s(x)) \times d\, e^{-ik\,x\cdot d} ds(x) \right) ds(d). \end{aligned} \tag{4.11}$$

Equations (4.8), (4.10), and (4.11) now imply the lemma.

We now define the *electric far field operator* $F : T^2(\Omega) \to T^2(\Omega)$ corresponding to the electric far field pattern E_∞ by

$$(Fg)(\hat{x}) := \int_\Omega E_\infty(\hat{x}; d, g(d))\, ds(d) \tag{4.12}$$

and note that F is a compact operator on $T^2(\Omega)$. The connection between the electric far field operator and electromagnetic Herglotz pairs is given by the following theorem.

THEOREM 4.3. *Let E_g^i, H_g^i and E_h^i, H_h^i be electromagnetic Herglotz pairs with kernels $g, h \in T^2(\Omega)$, respectively and let E_g and E_h be the solutions of* (4.1)–(4.3) *with E^i, H^i replaced by E_g^i, H_g^i and E_h^i, H_h^i, respectively. Then*

$$k \int\int_B \operatorname{Im} n\, E_g \cdot \overline{E_h}\, dx = -2\pi(Fg, h) - 2\pi(g, Fh) - (Fg, Fh)$$

where $(\cdot\,,\cdot)$ denotes the inner product on $T^2(\Omega)$.

Proof. Let E_g^s, E_h^s denote the scattered electric fields corresponding to E_g, H_g and E_h, H_h, respectively and let $E_{g,\infty}, E_{h,\infty}$ be the corresponding electric far field patterns. Then, using the vector Green's theorem, Lemma 4.1, and Lemma 4.2 we have

$$
\begin{aligned}
(4.13)\quad 2ik^2 \int\int_B \operatorname{Im} n\, E_g \cdot \overline{E_h}\, dx &= \int_{\partial B} \left(\nu \times E_g \cdot \operatorname{curl} \overline{E_h} - \nu \times \overline{E_h} \cdot \operatorname{curl} E_g\right) ds \\
&= \int_{\partial B} \left(\nu \times E_g^s \cdot \operatorname{curl} \overline{E_h^s} - \nu \times \overline{E_h^s} \cdot \operatorname{curl} E_g^s\right) ds \\
&+ \int_{\partial B} \left(\nu \times E_g^s \cdot \operatorname{curl} \overline{E_h^i} - \nu \times \overline{E_h^i} \cdot \operatorname{curl} E_g^s\right) ds \\
&+ \int_{\partial B} \left(\nu \times E_g^i \cdot \operatorname{curl} \overline{E_h^s} - \nu \times \overline{E_h^s} \cdot \operatorname{curl} E_g^i\right) ds \\
&= -2ik \int_\Omega E_{g,\infty} \cdot \overline{E_{h,\infty}}\, ds - 4\pi ik \int_\Omega E_{g,\infty} \cdot \overline{h}\, ds - 4\pi ik \int_\Omega g \cdot \overline{E_{h,\infty}}\, ds \\
&= -2ik(Fg, Fh) - 4\pi ik(Fg, h) - 4\pi ik(g, Fh)\,,
\end{aligned}
$$

noting that Fg is the electric far field pattern corresponding to the incident electric field E_g^i.

We note that the different factors ik^2 and k occurring in Theorems 2.3 and 4.3 are due solely to our normalization of the Herglotz wave functions and electromagnetic Herglotz pairs and have no mathematical or physical significance.

COROLLARY 4.4. *Assume that* $\operatorname{Im} n \geq 0$. *Then, except for possibly zero, F has no purely imaginary eigenvalues.*

Proof. Suppose $Fg = \lambda g$ with $\operatorname{Re} \lambda = 0, \lambda \neq 0$. Then, choosing $h = g$ in Theorem 4.3, we have

$$
\begin{aligned}
(4.14)\quad k \int\int_B \operatorname{Im} n\, |E_g|^2 dx &= -4\pi \operatorname{Re}(Fg, g) - (Fg, Fg) \\
&= -4\pi \operatorname{Re} \lambda \|g\|^2 - \|Fg\|^2.
\end{aligned}
$$

Since $\operatorname{Re} \lambda = 0$ we now have that $Fg = 0$ and hence, since $\lambda \neq 0, g = 0$.

COROLLARY 4.5. *Assume that* $\operatorname{Im} n = 0$. *Then F is normal and hence has a countable number of eigenvalues.*

Proof. From Theorem 4.3 we have that

$$
(4.15)\quad (Fg, Fh) = -2\pi\{(Fg, h) + (g, Fh)\}
$$

and hence

$$
(4.16)\quad (g, F^*Fh) = -2\pi\{(g, Fh) + (g, F^*h)\}
$$

for all $g, h \in T^2(\Omega)$. We can now conclude that

$$
(4.17)\quad F^*F = -2\pi\{F + F^*\}.
$$

We now note that by reciprocity [3, p. 179] we have

$$
\begin{aligned}
(4.18)\quad (Fh, g) &= \int_\Omega \int_\Omega E_\infty(d; \hat{x}, h(\hat{x})) \cdot \overline{g(d)}\, ds(\hat{x})\, ds(d) \\
&= \int_\Omega \int_\Omega h(\hat{x}) \cdot E_\infty(-\hat{x}; -d, \overline{g(d)})\, ds(d)\, ds(\hat{x})
\end{aligned}
$$

for all $g, h \in T^2(\Omega)$ and hence

$$(F^*g)(\hat{x}) = \int_\Omega \overline{E_\infty(-\hat{x}; -d, \overline{g(d)})}\, ds(d) \quad . \tag{4.19}$$

If we now define $R : T^2(\Omega) \to T^2(\Omega)$ by $(Rg)(d) := g(-d)$ we have that

$$F^*g = \overline{RFR\bar{g}}. \tag{4.20}$$

Now proceeding as in the proof of Corollary 2.5 we find that

$$FF^* = -2\pi\{F + F^*\} \tag{4.21}$$

and the proof is finished.

5. The inverse scattering problem for electromagnetic waves in an inhomogeneous medium. As in the scalar case, the analysis of the preceding section suggests a new method for solving the inverse scattering problem for electromagnetic waves in an inhomogeneous medium. In particular, let Γ be a smooth surface contained in the interior of the support of m such that k is not a Maxwell eigenvalue [3, p. 168] for the interior of Γ. For $\varphi \in T^2(\Gamma)$, define the vector potential

$$(A_\varphi)(x) := \int_\Gamma \Phi(x, y)\varphi(y)\, ds(y), \quad x \in \mathbb{R}^3 \setminus \Gamma, \tag{5.1}$$

and corresponding electromagnetic field

$$\tilde{E}_\varphi := \operatorname{curl} A_\varphi, \quad \tilde{H}_\varphi := \frac{1}{ik} \operatorname{curl} \tilde{E}_\varphi \tag{5.2}$$

and note that the electric far field pattern $\tilde{E}_{\varphi,\infty}$ of $\tilde{E}_\varphi$ is given by

$$\tilde{E}_{\varphi,\infty}(\hat{x}) = \frac{ik}{4\pi}\, \hat{x} \times \int_\Gamma e^{-ik\hat{x}\cdot y}\varphi(y)\, ds(y). \tag{5.3}$$

Finally, for $\lambda \neq 0$ define $g \in T^2(\Omega)$ by

$$\lambda g(\hat{x}) = \tilde{E}_{\varphi,\infty}(\hat{x}) \tag{5.4}$$

and the electromagnetic Herglotz pair E_g, H_g by

$$E_g(x) = \int_\Omega E^i(x; d, g(d))\, ds(d), \quad H_g(x) = \int_\Omega H^i(x; d, g(d))\, ds(d). \tag{5.5}$$

Now let E_∞ be the electric far field pattern corresponding to (4.1)–(4.3) and F the far field operator corresponding to E_∞. For $\lambda \notin \sigma(F)$, let $g_\ell^{(i)} \in T^2(\Omega)$ be the unique solution of

$$Fg - \lambda g = E_{\infty,\ell}^{(i)}, \tag{5.6}$$

where $\ell = 0, 1, 2, \ldots$, $i = 1, 2$, and

$$\begin{aligned} E_{\infty,\ell}^{(1)} &:= \frac{1}{i^\ell} \operatorname{Grad} Y_\ell, \\ E_{\infty,\ell}^{(2)} &:= -\frac{1}{ki^{\ell+1}} \nu \times \operatorname{Grad} Y_\ell, \end{aligned} \tag{5.7}$$

with Grad denoting the surface gradient. Then, since the scattered electric field is uniquely determined by its far field pattern [3, p. 157], we have from (5.6) and unique continuation that if there exists a unique solution $\varphi_\ell^{(i)}$ to the integral equation $\lambda g_\ell^{(i)} = E_{\varphi,\infty}$ (c.f. [3, p. 197]) then

$$\int_\Omega E(x;d,g_\ell^{(i)}(d))\,ds(d) - Eg_\ell^{(i)}(x) - \tilde{E}_{\varphi_\ell^{(i)}}(x) = E_\ell^{(i)}(x) \tag{5.8}$$

for $x \in \mathbb{R}^3 \setminus B$, where

$$\begin{aligned} E_\ell^{(1)}(x) &:= \operatorname{curl}\operatorname{curl}\{x h_\ell^{(1)}(k|x|)Y_\ell(\hat{x})\}, \\ E_\ell^{(2)}(x) &:= \operatorname{curl}\{x h_\ell^{(1)}(k|x|)Y_\ell(\hat{x})\}\,. \end{aligned} \tag{5.9}$$

As in the scalar case, one now tries to solve the inverse scattering problem of determining n from E_∞ by trying to construct n from the overdetermined system (5.6)–(5.9). If this problem is formulated as an optimization problem, then, as in the scalar case, it is easily seen that the minimum value of the cost functional will be zero provided the following theorem is true, where

$$\begin{aligned} Y := \{&(\mathcal{E},\mathcal{H}) \in C^1(\bar{B}) \times C^1(\bar{B}) : \operatorname{curl}\mathcal{E} - ik\mathcal{H} = 0, \\ &\operatorname{curl}\mathcal{H} + ikn(x)\mathcal{E} = 0 \text{ in } B\}, \\ W := \{&\nu \times (\mathcal{E} - E_g - \tilde{E}_\varphi),\ \nu \times \operatorname{curl}(\mathcal{E} - E_g - \tilde{E}_\varphi) : \\ &\varphi \in T^2(\Gamma),\ \lambda g = \tilde{E}_{\varphi,\infty}\,, (\mathcal{E},\mathcal{H}) \in Y\}, \end{aligned}$$

and B is as defined in the previous sections.

THEOREM 5.1. *Suppose $ik\lambda \notin \sigma(F)$. Then W is dense in $T^2(\partial B) \times T^2(\partial B)$.*

Proof. We first note that from (3.8) we have

$$\begin{aligned} E_g(x) &= ik \int_\Omega e^{ik\,x\cdot d}\, g(d)\,ds(d) \\ &= \frac{1}{4\pi\lambda} \operatorname{curl} \int_\Omega \int_\Gamma \varphi(y) e^{ik\,d\cdot(x-y)}\,ds(y)\,ds(d) \\ &= \frac{1}{\lambda} \operatorname{curl} \int_\Gamma \varphi(y) j_0(k|x-y|)\,ds(y). \end{aligned} \tag{5.10}$$

Now suppose that $a, b \in T^2(\partial B)$ satisfy

$$\int_{\partial B} \{(\mathcal{E} - E_g - \tilde{E}_\varphi)\cdot \bar{a} + \operatorname{curl}(\mathcal{E} - E_g - \tilde{E}_\varphi)\cdot \bar{b}\} ds = 0 \tag{5.11}$$

for all $(\mathcal{E},\mathcal{H}) \in Y$ and $\varphi \in T^2(\Gamma)$. Define

$$A(x) := \frac{1}{k^2} \operatorname{curl}\operatorname{curl} \int_{\partial B} \Psi(x,y)\overline{a(y)}\,ds(y) + \operatorname{curl} \int_{\partial B} \Psi(x,y)\overline{b(y)}\,ds(y) \tag{5.12}$$

for $x \in \mathbb{R}^3 \setminus \partial B$, where Ψ is given by (3.12). Then $E := \operatorname{curl} A$ is given by

$$E(x) = \operatorname{curl} \int_{\partial B} \Psi(x,y)\overline{a(y)}\,ds(y) + \operatorname{curl}\operatorname{curl} \int_{\partial B} \Psi(x,y)\overline{b(y)}\,ds(y) \tag{5.13}$$

for $x \in \mathbb{R}^3 \setminus \partial B$ and hence $\operatorname{curl} E = k^2 A$ for $x \in \mathbb{R}^3 \setminus \partial B$. Setting $\mathcal{E} = 0$ in (5.11) and using the vector identities

$$\begin{aligned} &\operatorname{curl}_y\{\Psi(x,y)\varphi(x)\} \cdot \overline{a(y)} = \varphi(x) \cdot \operatorname{curl}_x\{\Psi(x,y)\overline{a(y)}\}, \\ &\operatorname{curl}_y \operatorname{curl}_y\{\Psi(x,y)\varphi(x)\} \cdot \overline{b(y)} = \varphi(x) \cdot \operatorname{curl}_x \operatorname{curl}_x\{\Psi(x,y)\overline{b(y)}\} \end{aligned} \tag{5.14}$$

now shows that

$$\int_\Gamma \varphi \cdot E \, ds = 0 \tag{5.15}$$

for all $\varphi \in T^2(\Omega)$. Hence $\nu \times E = 0$ on Γ and since k is not a Maxwell eigenvalue for the interior of Γ we have that $E = 0$ in the interior of Γ. By unique continuation, $E = 0$ in B and hence $A = 0$ in B also. The L^2–jump relations for vector potentials [3, p. 165] now imply that

$$\nu \times \operatorname{curl} A_+ = \bar{a}, \quad \nu \times A_+ = \bar{b} \quad \text{on } \partial B \tag{5.16}$$

where $+$ again denotes the limit taken from the exterior of B. Setting $\varphi = 0$ in (5.11) now implies that

$$\int_{\partial B} \{\nu \times \mathcal{E} \cdot \operatorname{curl} A_+ - \nu \times A_+ \cdot \operatorname{curl} \mathcal{E}\} \, ds = 0 \tag{5.17}$$

for all $(\mathcal{E}, \mathcal{H}) \in Y$.

We now define

$$A^s(x) := \frac{1}{k^2} \operatorname{curl}\operatorname{curl} \int_{\partial B} \Phi(x,y)\overline{a(y)} \, ds(y) + \operatorname{curl} \int_{\partial B} \Phi(x,y)\overline{b(y)} \, ds(y) \tag{5.18}$$

for $x \in \mathbb{R}^3 \setminus \bar{B}$ and

$$\begin{aligned} A^0(x) := &\frac{1}{\lambda k^2} \operatorname{curl}\operatorname{curl} \int_{\partial B} j_0(k|x-y|)\overline{a(y)} \, ds(y) \\ &+ \frac{1}{\lambda} \operatorname{curl} \int_{\partial B} j_0(k|x-y|)\overline{b(y)} \, ds(y) \end{aligned} \tag{5.19}$$

for $x \in \mathbb{R}^3$ and note that from (3.8) we have

$$A^0(x) = \int_\Omega e^{ik\,x\cdot d} \, \tilde{g}(d) \, ds(d), \tag{5.20}$$

where

$$\tilde{g}(d) := \frac{1}{4\pi\lambda} \int_{\partial B} e^{-ik\,y\cdot d} \{(d \times \overline{a(y)}) \times d + ik\, d \times \overline{b(y)}\} \, ds(y). \tag{5.21}$$

In particular, from (5.10) and (5.21) we see that the far field pattern A_∞ of A^s is given by $A_\infty = \lambda \tilde{g}$.

In (5.17) we now set

$$\begin{aligned} \mathcal{E}(x) &= E^i(x; -\hat{z}, p) + E^s(x; -\hat{z}, p), \\ \mathcal{H}(x) &= H^i(x; -\hat{z}, p) + H^s(x; -\hat{z}, p), \end{aligned} \tag{5.22}$$

where $\hat{z} \in \Omega, p \in \mathbb{R}^3$, and use the vector Green's theorem to deduce that

$$\begin{aligned}(5.23)\qquad &\int_{\partial B} \{\nu(x) \times E^i(x;-\hat{z},p) \cdot \operatorname{curl} A^s(x) \\ &\qquad -\nu(x) \times A^s(x) \cdot \operatorname{curl} E^i(x;-\hat{z},p)\}\, ds(x) \\ &+ \int_{\partial B} \{\nu(x) \times E^s(x;-\hat{z},p) \cdot \operatorname{curl} A^0(x) \\ &\qquad -\nu(x) \times A^0(x) \cdot \operatorname{curl} E^s(x;-\hat{z},p)\}\, ds(x) = 0.\end{aligned}$$

Setting $ikB^s := \operatorname{curl} A^s$ and using (4.8) we now see that the first integral in (5.23) is given by

$$\begin{aligned}(5.24)\qquad &\int_{\partial B} \{\nu(x) \times E^i(x;-\hat{z},p) \cdot \operatorname{curl} A^s(x) \\ &\qquad -\nu(x) \times A^s(x) \cdot \operatorname{curl} E^i(x;-\hat{z},p)\}\, ds(x) \\ &= -ik \int_{\partial B} \{E^i(x;-\hat{z},p) \cdot \nu(x) \times B^s(x) \\ &\qquad +H^i(x;-\hat{z},p) \cdot \nu(x) \times A^s(x)\}\, ds(x) \\ &= k^2\, p \cdot \left(\hat{z} \times \int_{\partial B} \{\nu(x) \times A^s(x) + (\nu(x) \times B^s(x)) \times \hat{z}\} e^{-ik\,x\cdot\hat{z}}\, ds(x) \right) \\ &= -4\pi i k\, p \cdot A_\infty(\hat{z}).\end{aligned}$$

Inserting (5.20) into the second integral in (5.23) and interchanging orders of integration, we now have from (4.8) and reciprocity that

$$\begin{aligned}(5.25)\qquad &\int_{\partial B} \{\nu(x) \times E^s(x;-\hat{z},p) \cdot \operatorname{curl} A^0(x) - \\ &\qquad -\nu(x) \times A^0(x) \cdot \operatorname{curl} E^s(x;-\hat{z},p)\}\, ds \\ &= ik \int_\Omega \tilde{g}(d) \cdot \left(d \times \int_{\partial B} \{-\nu(x) \times E^s(x) + \right. \\ &\qquad \left. + [\nu(x) \times H^s(x)] \times d\} e^{ik\,x\cdot d}\, ds(x) \right) ds(d) \\ &= 4\pi\, p \cdot \int_\Omega E_\infty(\hat{z}; d, \tilde{g}(d))\, ds(d) \\ &= 4\pi p \cdot (F\tilde{g})(\hat{z}) \ .\end{aligned}$$

From (5.23)–(5.25) we now have that

$$(5.26)\qquad F\tilde{g} = ik\, A_\infty = ik\, \lambda \tilde{g}.$$

Since $ik\,\lambda \notin \sigma(F)$, we have that $\tilde{g} = 0$. This implies $A_\infty = 0$ and hence $A^s = 0$ in $\mathbb{R}^3 \setminus \bar{B}$ and $A^0 = 0$ in $\mathbb{R}^3$. Hence, $A = 0$ in $\mathbb{R}^3 \setminus \bar{B}$ and from (5.16) we now have that $a = b = 0$. The theorem is now proved.

REFERENCES

[1] S. K. Cho, *Electromagnetic Scattering*, Springer-Verlag, New York, 1990.

[2] D. COLTON AND P. HÄHNER, *Modified far field operators in inverse scattering theory*, SIAM J. Math. Anal., 24 (1993), pp. 365–389.
[3] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1992.
[4] D. COLTON AND P. MONK, *On a class of integral equations of the first kind in inverse scattering theory*, SIAM J. Appl. Math., 53 (1993), pp. 847–860.
[5] P.D. LAX AND R.S. PHILLIPS, *Scattering Theory*, Academic Press, New York, 1967.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 616–632, May 1995

006

THE INVERSE EIGENVALUE PROBLEM WITH FINITE DATA FOR PARTIAL DIFFERENTIAL EQUATIONS*

DAVID C. BARNES† AND ROGER KNOBEL‡

Abstract. This work is concerned with the inverse eigenvalue problem for the partial differential equation $\nabla^2 u + (\lambda - q(x,y))u = 0$. We study the problem of reconstructing the coefficient function $q(x,y)$ (or at least a numerical approximation to it) using only a finite amount of spectral data, say, $\lambda_n(q)$ for $n = 1, 2, \ldots, N$. One of the essential tasks considered here is that of determining how much information about the unknown function can be contained in such a fixed and finite amount of spectral data. A numerical method, based on a constrained least squares procedure, is devised for extracting such information, and several examples are given. A proof of convergence for the numerical method is provided. We show that the main difficulty with the finite inverse problem is that the eigenvalues are continuous in some very weak topologies. This work is a higher-dimensional version of the problem considered by Barnes [*SIAM J. Math Anal.*, 22 (1991), pp. 732–753] for ordinary differential equations.

Key words. inverse eigenvalue problem, continuous dependence

AMS subject classification. 35B25

1. Formulation of the finite inverse problem.

1.1. Introduction. Assume that $\mathcal{D}$ is a bounded piecewise smooth domain in the x, y plane, let $\vec{\nu}$ be a unit normal vector on $\partial\mathcal{D}$, and consider the eigenvalue problem

$$\nabla^2 u + (\lambda - q(x,y))u = 0, \quad \text{with} \quad \sigma_1 u + \sigma_2 \frac{\partial u}{\partial \vec{\nu}} = 0 \quad \text{on} \quad \partial\mathcal{D}. \tag{1}$$

Here, σ_1 and σ_2 are given functions, but both cannot vanish at the same time. Denote the eigenvalues of (1) by $\lambda_n(q)$ and let $q^*(x,y)$ represent the unknown coefficient function. Suppose that a finite amount of spectral data $\boldsymbol{\Lambda} = (\Lambda_1, \Lambda_2, \ldots, \Lambda_N)$ is given so that $\lambda_n(q^*) = \Lambda_n$ for $n = 1, 2, \ldots, N$. We want to construct the best possible approximation to $q^*(x,y)$ using the given data. A problem similar to this has been considered by Seidman [13]; however, he assumed that the domain $\mathcal{D}$ was the unit ball in N-dimensional space and that the coefficient functions were all radially symmetric. Such symmetry reduces the higher-dimensional problem to the one-dimensional case. Our work does not make such assumptions. For some additional results concerning this problem, see [3]–[5], [10], and [12].

Clearly, such a finite inverse problem cannot be solved uniquely. Therefore, we understand that a solution to such a problem is simply a sequence of functions $q_N(x,y)$ which has the correct spectral behavior—that is, a sequence satisfying the following interpolation condition: For each N, there is an ϵ_N such that

$$|\lambda_i(q_N) - \lambda_i(q^*)| < \epsilon_N \quad \text{for } i = 1, 2, \ldots, N \quad \text{and} \quad \epsilon_N \to 0 \quad \text{as } N \to \infty. \tag{2}$$

Next, we must provide a proper mathematical foundation for understanding the results. In particular, it is critical that the sequence $q_N(x,y)$ should converge to $q^*(x,y)$

* Received by the editors August 2, 1993; accepted for publication October 14, 1993.
† Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington, 99164-3113 (barnes@alpha.math.wsu.edu).
‡ Department of Mathematics and Computing Science, University of Texas–Pan American, Edinburgh, Texas 78539-2999.

in some suitable topology as $N \to \infty$. If so, it is necessary to understand something about the topology used and how accurate the approximation is. Also, for a given fixed value of N, we need to have some idea of how many different functions (in addition to q^*) might satisfy the spectral conditions $\boldsymbol{\lambda}(q) = \boldsymbol{\Lambda}$.

Theorem 2.4 and Corollaries 2.3 and 2.7 provide some partial answers to these questions. They show that, assuming a uniqueness condition, as $N \to \infty$, the approximating sequence $q_N(x, y)$ converges to $q^*(x, y)$ in a certain norm (we call it the 1Max norm) and that the eigenvalues are continuous in the topology generated by it. The continuity theorem indicates what kind of variation in the solution of such an inverse problem is still possible, even assuming that the uniqueness condition is satisfied and that all of the requirements of a finite amount of spectral data have been met. This analysis provides a proper mathematical foundation for the finite inverse problem.

A useful tool is the following variational characterization of eigenvalues using the Rayleigh quotient [6], [14]. Let $\mathcal{L}(\cdot)$ be a self-adjoint operator defined on a dense subspace $\mathbf{D}$ of a separable Hilbert space. Suppose that the lower part of the spectrum of $\mathcal{L}(\cdot)$ consists of isolated eigenvalues $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots$, each with finite multiplicity. Let u_i denote the eigenfunction corresponding to λ_i and let $\mathcal{U}_n$ be the subspace spanned by the first n eigenfunctions $u_1, u_2, \ldots, u_n$. Let $\mathcal{V}_n$ be any other n-dimensional subspace of $\mathbf{D}$. Let $\mathcal{R}(u)$ denote the Rayleigh quotient for $\mathcal{L}(\cdot)$. It is defined by

$$\mathcal{R}(u) = \frac{(\mathcal{L}(u), u)}{(u, u)}.$$

It follows that

$$\lambda_n \leq \max_{u \in \mathcal{V}_n} \mathcal{R}(u) \quad \text{and that} \quad \lambda_n = \max_{u \in \mathcal{U}_n} \mathcal{R}(u). \tag{3}$$

2. The topology of the finite inverse eigenvalue problem.

2.1. Continuity properties of the eigenvalues. In a bounded two-dimensional domain $\mathcal{D}$, let $\mathcal{C}(\mathcal{D}, H)$ be the class of measurable functions $q(x, y)$ with domain $\mathcal{D}$ satisfying $|q| \leq H$. It is convenient to assume that all functions in $\mathcal{C}(\mathcal{D}, H)$ vanish outside of $\mathcal{D}$. A major part of the theory of the finite inverse problem hinges on a topological analysis of the continuity properties of the eigenvalues $\lambda_n(q)$, which are simply real-valued functions defined on $\mathcal{C}(\mathcal{D}, H)$. To understand why, consider a classical theorem which says that the eigenvalues are continuous functions of q in the L_∞ norm. It is clearly not possible to construct any kind of reasonable approximation to the partial derivatives of $q(x, y)$ using only a finite amount of spectral data. But this continuity theorem does provide a little insight into the information content of the spectral data.

On the other hand, it follows easily from Corollary 2.3 that the eigenvalues are still continuous in the much weaker L_2 norm. Therefore, we cannot even construct a uniform approximation to $q^*(x, y)$. With each new and weaker topology in which the eigenvalues are continuous, we understand more about the information content of the spectral data because we know more about what kind of information that the data does *not* contain. This leads us to consider the problem of characterizing, in an understandable way, the weakest topology in which the first N eigenvalues are continuous.[1] It is this topology which provides an accurate measure of the information

[1] A basis for the open sets in this topology may be taken as the collection of subsets of $\mathcal{C}(\mathcal{D}, H)$ which are inverse images of an open interval under the mapping $\lambda_n : q \to \Re$ for $n = 1, 2, \ldots, N$.

content of the spectral data. Because the topology depends on N, it seems to be very difficult to characterize it exactly, but we will provide some approximations to it.

We will be especially concerned with the following three different topologies defined on $\mathcal{C}(\mathcal{D}, H)$: the topology of weak convergence of a sequence,[2] denoted by $q_n(x,y) \rightsquigarrow q(x,y)$, and the topologies defined by a pair of norms $\|\cdot\|_{1\text{Max}}$ and $\|\cdot\|_{\nabla^2}$. First, we will define the norms.

Given a function $q(x,y) \in \mathcal{C}(\mathcal{D}, H)$ and coefficients σ_1, σ_2 defined on $\partial\mathcal{D}$, let $\overrightarrow{\nu}$ be a unit normal vector and let $f(x,y)$ be the solution of the problem

$$\nabla^2 f = q(x,y) \quad \text{with} \quad 2\sigma_1 f + \sigma_2 \frac{\partial f}{\partial \overrightarrow{\nu}} = 0 \quad \text{on} \quad \partial\mathcal{D}. \tag{4}$$

Since an extra factor of 2 is included in the boundary condition of (4), the integral formula

$$\iint_{\mathcal{D}} q(s,t) u_n^2 \, ds\, dt = \iint_{\mathcal{D}} \left[\nabla^2 f(x,y)\right] u_n^2 \, ds\, dt = \iint_{\mathcal{D}} f(x,y) \left[\nabla^2 u_n^2\right] \, ds\, dt$$

will hold. It is easy to check that the following relation defines a norm $\|\cdot\|_{\nabla^2}$ on $\mathcal{C}(\mathcal{D}, H)$. For $q(x,y) \in \mathcal{C}(\mathcal{D}, H)$, let

$$\|q\|_{\nabla^2} = \max_{(x,y)\in\mathcal{D}} |f(x,y)|$$

with f defined by (4). Using Green's function $\mathcal{G}(x,y,s,t)$ for (4), it follows that

$$f(x,y) = \iint_{\mathcal{D}} \mathcal{G}(x,y,s,t) q(s,t) \, ds\, dt, \tag{5}$$

so that the ∇^2 norm also may be expressed as

$$\|q\|_{\nabla^2} = \max_{x,y\in\mathcal{D}} |f(x,y)| = \max_{x,y\in\mathcal{D}} \left| \iint_{\mathcal{D}} \mathcal{G}(x,y,s,t) q(s,t) \, ds\, dt \right|.$$

We will also be concerned with the 1Max norm defined by

$$\|q\|_{1Max} = \max_{x,y\in\mathcal{D}} \left| \int_0^x \int_0^y q(s,t) \, ds\, dt \right|.$$

We are assuming, for convenience, that the domain $\mathcal{D}$ is contained in the positive quadrant and that $q(x,y)$ vanishes outside $\mathcal{D}$. In analogy with the ∇^2 norm, the 1Max norm could also be defined as the maximum value of the function $f(x,y)$, the unique solution of the characteristic initial value problem $f_{xy} = q(x,y)$ with boundary conditions $f(0,y) = f(x,0) = 0$.

Now that these three topologies have been introduced, we will show that, in fact, all of them are equivalent in the class $\mathcal{C}(\mathcal{D}, H)$. Therefore, compactness of $\mathcal{C}(\mathcal{D}, H)$ in the weak topology gives compactness in both norm topologies. One reason to consider two different but equivalent norms is that it is difficult to compute $\|\cdot\|_{\nabla^2}$ since the Poisson equation (4) must be solved while it requires only a quadrature to compute

[2] Weak convergence means that $\int\int_{\mathcal{D}} f(x,y) q_n(x,y) \, dx\, dy \to \int\int_{\mathcal{D}} f(x,y) q(x,y) \, dx\, dy$ for every function $f(x,y)$ having $\int\int_{\mathcal{D}} f^2(x,y) \, dx\, dy < \infty$. The weak topology is useful because it makes $\mathcal{C}(\mathcal{D}, H)$ compact [6].

$\|\cdot\|_{1\text{Max}}$. However, it turns out that, at least for the study of inverse problems, it is analytically more convenient to use $\|\cdot\|_{\nabla^2}$. Furthermore, we will show later that, if $H\to\infty$, then the three topologies are not equivalent.

THEOREM 2.1. *If $q(x,y)$, $q_n(x,y) \in \mathcal{C}(\mathcal{D},H)$, then as $n\to\infty$, the following are all equivalent:*

(1) $q_n(x,y) \rightsquigarrow q(x,y)$,
(2) $\|q_n - q\|_{\nabla^2}\to 0$,
(3) $\|q_n - q\|_{1\text{Max}}\to 0$.

Proof. We will prove the logical implications (1) $\Leftrightarrow$ (2) and (1) $\Leftrightarrow$ (3). First, supposing that (1) holds, use Green's function to define $Q_n(x,y)$ by

$$Q_n(x,y) = \iint_{\mathcal{D}} \mathcal{G}(x,y,s,t)q_n(s,t)\,ds\,dt.$$

Let $\Delta q_n = q_n - q$ and $\Delta Q_n = Q_n - Q$. Then the weak convergence implies, for each fixed point (x,y), that

$$\Delta Q_n(x,y) = \iint_{\mathcal{D}} \mathcal{G}(x,y,s,t)\Delta q_n(s,t)\,ds\,dt\to 0.$$

However, $\Delta Q_n(x,y)$ is uniformly bounded and equicontinuous, implying that there is a uniformly convergent subsequence. Since the limit must be unique, every such subsequence must converge to the same function. Thus, the overall convergence is uniform. Using Green's function characterization of the ∇^2 norm (5) shows that $\|q_n - q\|_{\nabla^2}\to 0$.

Now suppose (2). Let $h(x,y) \in \mathcal{C}(\mathcal{D},H)$ be any function smooth enough so that, for some choice of $r(x,y) \in L_2[\mathcal{D}]$, it can be represented in the form

$$h(x,y) = \iint_{\mathcal{D}} \mathcal{G}(x,y,s,t)r(s,t)\,dt\,ds.$$

Consider

$$\begin{aligned}(h,\Delta q_n) &= \iint_{\mathcal{D}} h(x,y)\Delta q_n(x,y)\,dx\,dy \\ &= \iint_{\mathcal{D}} r(s,t)\left(\iint_{\mathcal{D}} \mathcal{G}(x,y,s,t)\Delta q_n(x,y)\,dx\,dy\right) ds\,dt \\ &= \iint_{\mathcal{D}} r(s,t)\Delta Q_n(s,t)\,ds\,dt = (r,\Delta Q_n).\end{aligned}$$

Since $\Delta Q_n(s,t)$ converges uniformly to zero, it follows that $(h,q_n)\to 0$, which implies weak convergence for smooth functions $h(x,y)$. Since such smooth functions are dense in $\mathcal{C}(\mathcal{D},H)$, condition (1) follows.

Now suppose (1). Given $(x,y) \in \mathcal{D}$, let $h(s,t) = 1$ for $s \le x$ and $t \le y$; otherwise, let $h(s,t) = 0$. Weak convergence implies the following pointwise convergence (in (x,y)):

$$(h,\Delta q_n) = \iint_{\mathcal{D}} h(s,t)\Delta q_n(s,t)\,ds\,dt = \int_{-\infty}^{x}\int_{-\infty}^{y} \Delta q_n(s,t)\,dt\,ds\to 0.$$

However, the sequence of functions $\int_{-\infty}^{x}\int_{-\infty}^{y} q_n(s,t)\,ds\,dt$ is uniformly bounded and equicontinuous so that it will converge uniformly. Thus, $\|\Delta q_n\|_{1\text{Max}}\to 0$.

Now suppose (3). It follows that, if $\mathcal{X}(s,t)$ is the characteristic function of any rectangle contained in $\mathcal{D}$, then $\iint_{\mathcal{D}} \mathcal{X}(s,t)\Delta q_n(s,t)\,ds\,dt \to 0$. Taking linear combinations of such characteristic functions shows that the same result holds for any step function defined on a rectangular grid. Since such functions are dense, it follows that $q_n(x,y) \rightsquigarrow q(x,y)$. □

Incidentally, this theorem shows that the topology generated by the ∇^2 norm is really independent of the terms σ_1 and σ_2 used in the boundary conditions of (4).

We will now show that the eigenvalues of (1) are continuous on $\mathcal{C}(\mathcal{D},H)$ with respect to these three topologies. Of course, since all three are actually identical, we will need only one proof.

THEOREM 2.2. *Let $\lambda_n(q)$ denote the nth eigenvalue of* (1) *corresponding to $q(x,y)$, and let $q_1, q_2 \in \mathcal{C}(\mathcal{D},H)$. Let μ_n denote the nth eigenvalue of* (1) *corresponding to the constant function $q(x,y) = H$ so that $\mu_n = \lambda_n(H)$. Then for all $n = 1,2,\ldots,N$,*

$$|\lambda_n(q_1) - \lambda_n(q_2)| \le 4N(H + |\mu_N|)\|q_1 - q_2\|_{\nabla^2}. \tag{6}$$

Proof. Let u be any function of the general form

$$u = \sum_{i=1}^{N} \alpha_i u_i \quad \text{where} \quad \sum_{i=1}^{N} \alpha_i^2 = 1 \quad \text{and} \quad \iint_{\mathcal{D}} u_i^2\,dA = 1,$$

where u_i are eigenfunctions corresponding to some coefficient function $p \in \mathcal{C}(\mathcal{D},H)$. We will show that any such function u satisfies the inequality

$$\left|\iint_{\mathcal{D}} u^2(q_1 - q_2)\,dA\right| \le 4N(H + |\mu_N|)\|q_1 - q_2\|_{\nabla^2}. \tag{7}$$

By assuming for a moment that this is true, we see that

$$\iint_{\mathcal{D}} u^2 q_1\,dA \le \iint_{\mathcal{D}} u^2 q_2\,dA + 4N(H + |\mu_N|)\|q_1 - q_2\|_{\nabla^2}.$$

The Rayleigh quotient for this problem is given by

$$\mathcal{R}(u,q) = \frac{\iint_{\mathcal{D}} u_x^2 + u_y^2 + q(x,y)u^2\,dA + \int_{\partial\mathcal{D}} u\partial u/\partial\vec{\nu}\,ds}{\iint_{\mathcal{D}} u^2\,dA}.$$

Equation (7) shows that $\mathcal{R}(u,q_1) \le \mathcal{R}(u,q_2) + 4N(H + |\mu_N|)\|\Delta q\|_{\nabla^2}$. Now take the maximum over all functions u in the space spanned by the first N eigenfunctions corresponding to the function q_2. It follows from the Min/Max principle (3) that $\lambda_n(q_1) \le \lambda_n(q_2) + 4N(H+|\mu_N|)\|\Delta q\|_{\nabla^2}$. Reversing the roles of q_1 and q_2 and repeating this argument gives the other half of (6) and proves the theorem. Now we need only to prove (7).

Let $u = \sum_{i=1}^{N} \alpha_i u_i$ with $\sum_{i=1}^{N} \alpha_i^2 = 1$ and suppose that f solves (4). Then

$$\begin{aligned}\iint_{\mathcal{D}} u^2\Delta q\,dA = \iint_{\mathcal{D}} u^2\nabla^2 f\,dA &= \iint_{\mathcal{D}} f\nabla^2 u^2\,dA \\ &= \sum_{i,j=0}^{N} \alpha_i\alpha_j \iint_{\mathcal{D}} f\nabla^2(u_i u_j)\,dA.\end{aligned} \tag{8}$$

Now u_i, u_j are eigenfunctions corresponding to some other function $p \in \mathcal{C}(\mathcal{D}, H)$ so that $\nabla^2 u_i = (p - \lambda_i(p))u_i$. Thus,

$$|\nabla^2(u_i u_j)| \leq |u_j u_i(p - \lambda_i(p))| + 2|\nabla u_i||\nabla u_j| + |u_i u_j(p - \lambda_j(p))|.$$

This inequality shows that the expression $|\iint_{\mathcal{D}} f \nabla^2 u_i u_j \, dA|$ is bounded above by the following term:

$$\|\Delta q\|_{\nabla^2} \iint_{\mathcal{D}} |u_j u_i(p - \lambda_i(p))| + 2|\nabla u_i||\nabla u_j| + |u_i u_j(p - \lambda_j(p))| \, dA. \tag{9}$$

Using the elementary inequality $|u_j u_i(p - \lambda_i(p))| \leq |u_i u_j(H + \mu_N)$ followed by the normalizing condition and Cauchy's inequality shows that the integral of the first and last terms in (9) are bounded by $H + \mu_N$. Combining the relation $\int\int |\nabla u_i|^2 = \lambda_i - \int\int q u_i^2 \, dA$ with a second application of Cauchy's inequality shows that the middle term of (9) is bounded by $(H + |\lambda_i|)^{1/2}(H + |\lambda_j|)^{1/2}$. Now, replace $|\lambda_i(p)|$ and $|\lambda_j(p)|$ with the larger value μ_N, add all three terms together, and substitute into (8). Finally, use Cauchy's inequality again and the normalizing condition to obtain (7). □

COROLLARY 2.3. *The eigenvalues $\lambda_n(q)$ are continuous on $\mathcal{C}(\mathcal{D}, H)$ with respect to weak convergence as well as with respect to the* 1Max *and the* ∇^2 *norms.*

Since the L_p norm is stronger than either the 1Max *or the* ∇^2 *norm, it also follows that the eigenvalues are continuous in the L_p norm for any $p \geq 1$.*

The second part of the corollary follows from the inequality

$$\|q\|_{1Max} = \max_{x,y \in \mathcal{D}} \left| \int_0^x \int_0^y q(s,t) \, ds \, dt \right| \leq \int_0^\infty \int_0^\infty |q(s,t)| \, ds \, dt \leq \|q\|_p.$$

The assumption that the functions q are uniformly bounded is crucial to the success of Theorems 2.2 and 2.4. All three topologies become very different if we let $H \to \infty$. For example, let $\mathcal{D}$ be the unit square, use Dirichlet boundary conditions, and take a perturbation in $q(x,y)$ of the form $\Delta q_{i,j}(x,y) = H \sin(i\pi x + j\pi y)$ where i, j are integers. A short calculation shows that $\|\Delta q_{i,j}\|_{1\text{Max}} = 3H/(ij\pi^2)$. Solving for f in the equation $\nabla^2 f = \Delta q_{i,j}$ gives

$$f(x,y) = \frac{-H}{(i^2 + j^2)\pi^2} \sin(i\pi x + j\pi y) \quad \text{so that} \quad \|\Delta q_{i,j}\|_{\nabla^2} = \frac{H}{(i^2 + j^2)\pi^2}.$$

Now, take $i = 1$ and $H = j$ to find that $\|\Delta q_{1,j}\|_{1\text{Max}} = 3$, while $\|\Delta q_{1,j}\|_{\nabla^2} \to 0$ as $j \to \infty$. In the L_2 norm, however, $\|q_{1,j}\|_{L^2} \to \infty$. Since the L_2 norms of any weakly convergent sequence must be uniformly bounded, $\Delta q_{i,j}(x,y)$ does not converge weakly. It is also easy to construct examples where $\|\Delta q_{i,j}\|_{1\text{Max}} \to 0$ without converging weakly.

Thus, the eigenvalues are continuous in some very weak topologies, and this is the root of the difficulty with the finite inverse eigenvalue problem. To illustrate further, suppose we are trying to find a numerical solution of such a problem for (1). Suppose that the unknown function is $q^*(x,y)$ and let $f^*(x,y)$ be the corresponding solution of (4). Then any other solution $f(x,y)$ of (4) that is a good uniform approximation to $f^*(x,y)$ will provide an equally valid solution to the finite inverse eigenvalue problem. This means that, at most, the only information contained in a finite amount of spectral data is a uniform approximation to $f^*(x,y)$. Therefore, reconstructing a pointwise approximation to $q^*(x,y)$ requires one to apply the operator $\nabla^2(\cdot)$ to the uniform approximation f, resulting in a very ill conditioned operation. Unfortunately, the fact is that a finite amount of spectral data simply cannot yield any better information about the function $q^*(x,y)$.

2.2. Some convergence theorems. Although the spectral data may not contain all of the information we want it to have, we will now show that, under a uniqueness condition, any sequence of interpolating functions (2) will converge to $q*$ in the weak topology as $N\to\infty$. The fundamental tool is the compactness of the class $\mathcal{C}(\mathcal{D}, H)$.

THEOREM 2.4. *Let Λ_n denote some spectral data for the problem* (1). *Suppose that the data is sufficient to insure that the infinite inverse eigenvalue problem always has a unique solution. That is, there is a unique function $q^* \in \mathcal{C}(\mathcal{D}, H)$ for which $\lambda_n(q^*) = \Lambda_n$ for all n. Let q_N be any sequence of functions which interpolates to the data $\lambda_n(q^*)$ as in* (2). *Then as $N\to\infty$,*

$$q_N \rightsquigarrow q^* \quad \text{and} \quad \|q_N - q^*\|_{1\text{Max}}\to 0 \quad \text{and} \quad \|q_N - q^*\|_{1\text{Max}}\to 0.$$

Proof. The proof follows the lines of the one given in [4]. Let q_N be the sequence of interpolation functions. Since $\mathcal{C}(\mathcal{D}, H)$ is compact, we may select a weakly convergent subsequence $q_{N_j} \rightsquigarrow \bar{q} \in \mathcal{C}(\mathcal{D}, H)$ as $j\to\infty$. The interpolation condition (2) shows that $\lambda_n(q^*) = \lambda_n(\bar{q})$ for all n so that $q^* = \bar{q}$. Thus, for any convergent subsequence, $q_{N_j} \rightsquigarrow q^*$ as $j\to\infty$ so that $q_N \rightsquigarrow q^*$ as $N\to\infty$. □

Of course, the difficulty with this theorem is that currently very little is known about the uniqueness of the inverse eigenvalue problem for partial differential equations, although progress has been made in the corresponding problem for ordinary differential equations [11]. In one of the few known uniqueness results dealing with the inverse problem for (1), Nachman, Sylvester, and Uhlmann [12] gave the following condition under which the infinite inverse problem will have unique solutions.

THEOREM 2.5 (Nachman et al.). *Consider* (1) *with Dirichlet boundary conditions. Suppose that $q_1(x, y)$ and $q_2(x, y)$ are two functions for which $\lambda_n(q_1) = \lambda_n(q_2)$ for all n and also that on the boundary of $\mathcal{D}$, the corresponding eigenfunctions, called $u_n^{[1]}$ and $u_n^{[2]}$, have equal normal derivatives*

$$\frac{\partial u_n^{[1]}}{\partial \nu} = \frac{\partial u_n^{[2]}}{\partial \nu}, \tag{10}$$

then $q_1 = q_2$.

Combining Theorems 2.4 and 2.5 yields the following theorem.

THEOREM 2.6. *Suppose that q^* and q_N are functions in $\mathcal{C}(\mathcal{D}, H)$ that satisfy $\lambda_n(q^*) = \lambda_N(q_N)$ for $n = 1, 2, \ldots, N$ and suppose that all of their corresponding eigenfunctions satisfy the normal derivative condition* (10). *Then as $N\to\infty$,*

$$q_N \rightsquigarrow q^* \quad \text{and} \quad \|q_N - q^*\|_{1\text{Max}}\to 0 \quad \text{and} \quad \|q_N - q^*\|_{1\text{Max}}\to 0. \tag{11}$$

Since $\mathcal{C}(\mathcal{D}, H)$ is compact, we can find a function, $\widehat{q} \in \mathcal{C}(\mathcal{D}, H)$, and a weakly convergent subsequence, $q_{N_i}\to\widehat{q}$, as $i\to\infty$. However, Theorem 2.1 shows that $\lambda_n(\widehat{q}) = \lambda_n(q^*)$ for all n. Theorem 2.6 implies that $q^* = \widehat{q}$. Since every convergent subsequence of q_N converges to the same q^*, it follows that q_N must converge to q^* as $N\to\infty$.

It is easy to see that an analogous theorem will continue to hold under any kind of condition on the spectral data or on the class $\mathcal{C}(\mathcal{D}, H)$ that gives uniqueness of the infinite inverse eigenvalue problem.

Still, it seems likely that many other inverse problems for (1) will not have unique solutions when the class of functions $\mathcal{C}(\mathcal{D}, H)$ is unrestricted. It seems more reasonable to expect a uniqueness result if the class $\mathcal{C}(\mathcal{D}, H)$ is constrained in some way—say, by requiring the functions $q(x, y)$ to be convex or to have some kind of symmetry.

Frequently, in applied inverse problems, a great deal is known about the qualitative behavior of the unknown function. An elementary example of such a constraint is used in the numerical calculation given in §5 below.

As another use of the idea of a subclass of $\mathcal{C}(\mathcal{D}, H)$, consider the case in which it is known that the function q^* satisfies a uniform Lipschitz condition

$$|q^*(x,y) - q^*(s,t)| \leq L(|x-y| + |s-t|). \tag{12}$$

Thus, we may also constrain the interpolating functions q_N to satisfy (12) and still be assured of obtaining a solution of the inverse problem. Then weak convergence implies uniform convergence so that q_N converges uniformly to q. Even if the uniqueness condition is not satisfied, then compactness still assures us of the existence of a uniformly convergent subsequence. If the Lipschitz condition is assumed to hold only on some subset of the domain $\mathcal{D}$, then we may select the interpolating sequence q_N to satisfy the Lipschitz condition there; thus we are assured of uniform convergence on the subset and, at least, of weak convergence elsewhere. We will show below how such constraints on the interpolating sequence q_N may be numerically implemented.

2.3. The isospectral equivalence classes $\mathcal{C}_N(\mathcal{D}, H)$ of $\mathcal{C}(\mathcal{D}, H)$. Since uniqueness is not readily available, we will consider the behavior of the set of isospectral equivalence classes on $\mathcal{C}(\mathcal{D}, H)$. We assume that only a finite amount of data is given. Thus, two functions q_1, q_2 are equivalent if $\lambda_n(q_1) = \lambda_n(q_2)$ for $n = 1, 2, 3, \ldots, N$. Call this set of equivalence classes $\mathcal{C}_N(\mathcal{D}, H)$, which inherits a natural topology from $\mathcal{C}(\mathcal{D}, H)$. Reinterpreting Theorem 2.4 in the light of these equivalence classes provides the following result.

COROLLARY 2.7. *The infinite inverse eigenvalue problem has a unique solution if and only if the diameter (as measured in the* 1Max *norm) of each equivalence class $\mathcal{C}_N(\mathcal{D}, H)$ tends to zero as $N \to \infty$.*

Let R^N be Euclidean N-dimensional space and define a mapping

$$\Phi : \mathcal{C}(\mathcal{D}, H) \to R^N \quad \text{by} \quad \Phi(q) = \boldsymbol{\lambda}(q),$$

using the brief notation $\boldsymbol{\lambda}(q) = (\lambda_1(q), \lambda_2(q), \ldots, \lambda_N(q))$. Let $\mathcal{S}(\mathcal{D}, H) \subset R^N$ be the range of Φ. Define a topology on $\mathcal{S}(\mathcal{D}, H)$ by using the component-wise convergence criteria $\boldsymbol{\lambda}(q_j) \to \boldsymbol{\lambda}(q^*)$ if and only if, for each $n = 1, 2, \ldots, N$, $\lambda_n(q_j) \to \lambda_n(q^*)$ as $j \to \infty$. Now topologize $\mathcal{C}(\mathcal{D}, H)$, using the weakest topology in which the first N eigenvalues are continuous. This simply means that $q_j \to q^*$ as $j \to \infty$ if and only if $\lambda_n(q_j) \to \lambda_n(q^*)$ for each $n = 1, 2, \ldots, N$. Now Φ induces a natural mapping of the equivalence classes $\mathcal{C}_N(\mathcal{D}, H)$ onto $\mathcal{S}(\mathcal{D}, H) \subset R^N$, and the map will be both continuous and one-to-one. Thus, $\mathcal{C}_N(\mathcal{D}, H)$ is homeomorphic to a subset of ordinary N-dimensional Euclidean space. This determines the weakest topology on $\mathcal{C}_N(\mathcal{D}, H)$, in which the eigenvalues are continuous. Unfortunately, knowing the topology on $\mathcal{C}_N(\mathcal{D}, H)$ still does not easily translate into usable conditions on the coefficient functions q themselves which, of course, belong to $\mathcal{C}(\mathcal{D}, H)$ and not to $\mathcal{C}_N(\mathcal{D}, H)$.

3. Higher-order equations in many dimensions. Similar methods can be applied to more general eigenvalue problems. We will use the standard notation and terminology for general partial differential operators as given, for example, by Friedman [7].

Let Ω be a bounded set in p-dimensional space, let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ be a multi-index, and let $D^\alpha = D_1^{\alpha_1} D_2^{\alpha_2} \cdots D_p^{\alpha_p}$ where $D_j = \partial / \partial x_j$. Let $\|u\|_m^\Omega$ be the Sobolev norm of order m for functions u defined on Ω. Let $\mathcal{L}$ be the differential

operator defined by $\mathcal{L} = \sum_{|\alpha| \leq 2m} a_\alpha(\boldsymbol{x}) D^\alpha$ where $\boldsymbol{x} \in \Omega$. Details of the conditions on Ω and $\mathcal{L}$, under which eigenvalue problems for such operators are well posed, have been studied extensively [7] and will not be repeated here. We will simply assume that the operator is such that the variational principle (3) is valid and that the following standard a priori inequality will hold for all u in an appropriate Sobolev space [1], [7]:

$$\|u\|_{2m}^{\Omega} \leq C \left(\|\mathcal{L}(u)\|_0^{\Omega} + \|u\|_0^{\Omega} \right).$$

The constant C is independent of u. This inequality takes the place of those used in §2.2.

As an example of what can be done with these methods, we will study the eigenvalue problem

$$\mathcal{L}(u) + (\lambda - q(\boldsymbol{x}))u = 0 \quad \text{with} \quad \frac{\partial^j u}{\partial \mu^j} = 0 \quad \text{on} \quad \partial\Omega \quad \text{for} \quad j = 0, 1, 2 \ldots m-1. \tag{13}$$

Here, μ is the outward normal to the boundary. To define the appropriate norm of a function $q(\boldsymbol{x})$, let f be the unique solution of the Dirichlet problem

$$\mathcal{L}(f) = q(\boldsymbol{x}) \quad \text{with} \quad \frac{\partial^j f}{\partial \mu^j} = 0 \quad \text{on} \quad \partial\Omega \quad \text{for} \quad j = 0, 1, 2 \ldots m-1.$$

Now define the norm $\|q\|_{\mathcal{L}}$ by

$$\|q\|_{\mathcal{L}} = \max_{\boldsymbol{x} \in \Omega} |f(\boldsymbol{x})|.$$

The following theorem gives the continuity of the eigenvalues of (13).

THEOREM 3.1. *Let $\lambda_n(q)$ denote the nth eigenvalue of the problem* (13). *There is a constant $C_n(H, \mathcal{D})$ which depends only on H, n, and $\mathcal{D}$, for which*

$$|\lambda_n(q_1) - \lambda_n(q_2)| \leq C_n(H, \mathcal{D}) \|q_1 - q_2\|_{\mathcal{L}}.$$

Proof. Proceeding as before, we see that u^2 satisfies the boundary conditions (13), so that

$$\int_\Omega \Delta q u^2 \, dV = \int_\Omega \mathcal{L}(f) u^2 \, dV = \int_\Omega f \mathcal{L}(u^2) \, dV \leq \|f\|_\infty \int_\Omega |\mathcal{L}(u^2)| \, dV.$$

Suppose that the eigenfunction is normalized so that $\|u\|_0^{\Omega} = 1$. Now for any $u \in \mathcal{U}(q)$, the expression $\mathcal{L}(u^2)$ can be written as a linear combination of terms like $D^{\alpha_i} u_k D^{\alpha_j} u_l$. The Schwartz inequality shows that $\int_\Omega |\mathcal{L}(u^2)| \, dV$ is bounded by a sum of terms like $C \|u_k\|_i^{\Omega} \|u_l\|_j^{\Omega}$, where C is a constant. The rest of the proof is much like that for the two-dimensional case. □

It is now clear that all of the two-dimensional theorems have generalizations to these higher-dimensional, higher-order problems. It is also interesting to observe that the higher-order problems are more poorly conditioned than the lower-order ones. To justify this observation, consider the problem of attempting to obtain pointwise data about q from the spectral data $\boldsymbol{\Lambda}$. Theorem 2.6 shows that the only information about q that is contained in a finite amount of data is a uniform approximation to the solution f of $\mathcal{L}(f) = q(\boldsymbol{x})$. So reconstructing q requires that the differential operator $\mathcal{L}(\cdot)$ be applied to the uniform approximation to f, resulting in a very ill conditioned operation. Such a manipulation becomes more ill conditioned as the order of the operator increases.

4. Application to optimization problems. These continuity properties also have applications to certain variational methods for isoperimetric inequalities. One application is to the existence theory for "eigenvalue gap" problems like those considered in [2]. These problems ask for the extremals of quantities such as $\lambda_2(q) - \lambda_1(q)$ or $\lambda_2(q)/\lambda_1(q)$. Given our continuity theorems, it is easy to show the existence of extremals for such problems on $\mathcal{C}(\mathcal{D}, H)$ since the differences and ratios of eigenvalues are continuous functions on a compact set.

5. Numerical examples.

5.1. Introduction. In this section we present three examples illustrating the numerical solution of the finite data problem in two dimensions. Our approach here is a least squares method that generalizes the algorithm described in [4] for the inverse Sturm–Liouville problem. The first example uses Dirichlet eigenvalues to construct an approximation of a function $q(x, y)$ defined on a rectangle and containing certain symmetry properties. In the second example, the symmetry assumptions are dropped, and an approximation to a general $q(x, y)$ is accomplished, using eigenvalues from four sets of boundary conditions. Finally, our third example exhibits the approximation of a characteristic function $q(x, y)$.

In each of the examples, the given spectral data $\boldsymbol{\Lambda}$ consists of N eigenvalues of (1). Using the data $\boldsymbol{\Lambda}$, functions $q_{N,M} \in \mathcal{C}(\mathcal{D}, H)$ of the form

$$q_{N,M}(x, y) = \sum_{k=1}^{M} \beta_k h_k(x, y) \tag{14}$$

are constructed to approximate $q(x, y)$. In general, the number of basis functions M is assumed to be at least as large as the number of given eigenvalues N. In [4], for example, M was taken to be $N \leq M \leq 2N$, and here we generally use $M = N$. For $q_{N,M}(x, y)$ of the form (14), the resulting spectral data $\boldsymbol{\lambda}(q_{N,M})$ depends on the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^t$, and we denote it by $\boldsymbol{\lambda}(\boldsymbol{\beta})$. Our approximation $q_{N,M}(x, y)$ is then found by solving for the $\boldsymbol{\beta}$ that minimizes the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta})\|_2$.

Although any well-behaved basis could be used, our choice for $\{h_k(x, y)\}_{k=1}^{M}$ will be piecewise constant functions: subdividing $\mathcal{D}$ into disjoint subsets $\cup_{k=1}^{M} D_k$, we define the function $h_k(x, y)$ to be 1 on D_k and 0 otherwise. The condition $q_{N,M} \in \mathcal{C}(\mathcal{D}, H)$ is easily implemented by requiring $|\beta_k| \leq H$ for each k. As discussed earlier, other kinds of useful and interesting constraints may be imposed on the coefficients $\boldsymbol{\beta}$. An efficient way to implement such ideas numerically has been given by Hanson and Haskel [8]. The program can solve constrained least squares problems for matrices. Specifically, it solves the following problem:

$$\begin{aligned} &\text{Minimize} \quad \|\boldsymbol{C} - \Gamma\boldsymbol{\beta}\|_2 \quad \text{subject to constraints} \\ &E\boldsymbol{\beta} = \boldsymbol{F} \quad \text{and} \quad G\boldsymbol{\beta} \geq \boldsymbol{K}. \end{aligned} \tag{15}$$

Letting $\mathcal{B}(H)$ denote the set

$$\mathcal{B}(H) = \{\boldsymbol{\beta} \in \boldsymbol{R}^M \mid |\beta_k| \leq H\},$$

we will consider the nonlinear constrained least squares problem

$$\min_{\boldsymbol{\beta} \in \mathcal{B}(H)} \|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta})\|_2 = \|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta}^*)\|_2. \tag{16}$$

To minimize the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta})\|_2$, we approximate the nonlinear problem (16) with a sequence of linear least squares problems. Beginning with an initial guess $\boldsymbol{\beta}_0$, $\boldsymbol{\lambda}(\boldsymbol{\beta})$ in (17) is replaced by its linearization $L_0(\boldsymbol{\beta}) = \boldsymbol{\lambda}(\boldsymbol{\beta}_0) + \Gamma_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ about $\boldsymbol{\beta}_0$. Finding $\boldsymbol{\beta}_1 \in \mathcal{B}(H)$, which minimizes $\|\boldsymbol{\Lambda} - L_0(\boldsymbol{\beta})\|_2$ over the set $\mathcal{B}(H)$, constructs our new approximation. One may then repeat this procedure to produce a sequence of successive approximations by computing the linearization $L_i(\boldsymbol{\beta})$ of $\boldsymbol{\lambda}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_i$ and by finding a solution $\boldsymbol{\beta}_{i+1} \in \mathcal{B}(H)$ of

$$\min_{\boldsymbol{\beta} \in \mathcal{B}(H)} \|\boldsymbol{\Lambda} - L_i(\boldsymbol{\beta})\|_2 = \|\boldsymbol{\Lambda} - L_i(\boldsymbol{\beta}_{i+1})\|_2. \tag{17}$$

At this point, the least squares method of [8] may be used to implement any additional constraints on $\boldsymbol{\beta}_i$.

Much of the computational cost of the method lies in the construction of $L_i(\boldsymbol{\beta})$. To compute the linearization of the spectral data $\boldsymbol{\lambda}(\boldsymbol{\beta}) = (\lambda_1(\boldsymbol{\beta}), \ldots, \lambda_N(\boldsymbol{\beta}))^t$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_i$, one must compute eigenvalues and eigenfunctions of (1). In our numerical examples, we use the Rayleigh–Ritz method to estimate the eigenvalues $\boldsymbol{\lambda}(\boldsymbol{\beta})$. To compute the linear component Γ of the linearization, note that if $\lambda_k(\boldsymbol{\beta}_i)$ is a simple eigenvalue with L^2-normalized eigenfunction $\phi_k(x, y)$, then a formal calculation [9] of the derivative of $\lambda_k(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_i$ yields

$$\frac{\partial \lambda_k}{\partial \beta_j}(\boldsymbol{\beta}_i) = \iint_{\mathcal{D}} h_j(x, y) \phi_k^2(x, y) \, \mathrm{d}A = \iint_{\mathcal{D}_j} \phi_k^2(x, y) \, \mathrm{d}A.$$

The linearization $L_i(\boldsymbol{\beta}) = \boldsymbol{\lambda}(\boldsymbol{\beta}_i) + \Gamma_i(\boldsymbol{\beta} - \boldsymbol{\beta}_i)$ is then computed by forming the $N \times M$ matrix Γ_i given by

$$[\Gamma_i]_{jk} = \iint_{\mathcal{D}_j} \phi_k^2(x, y) \, \mathrm{d}A.$$

In our numerical calculations, we chose not to update Γ_i at each step, but rather to take Γ_0 as an approximation to Γ_i and to replace (17) with the minimization problem

$$\min_{\boldsymbol{\beta} \in \mathcal{B}(H)} \|\boldsymbol{\Lambda} - \tilde{L}_i(\boldsymbol{\beta})\|_2 = \|\boldsymbol{\Lambda} - \tilde{L}_i(\boldsymbol{\beta}_{i+1})\|_2, \tag{18}$$

where $\tilde{L}_i(\boldsymbol{\beta}) = \boldsymbol{\lambda}(\boldsymbol{\beta}) + \Gamma_0(\boldsymbol{\beta} - \boldsymbol{\beta}_i)$.

5.2. Recovery of a symmetric potential on a rectangle. Our first example is a generalization of the inverse Sturm–Liouville problem with symmetric potential. Here we consider a rectangular domain $\mathcal{D} = [0, \pi/a] \times [0, \pi]$ with functions $q(x, y)$ defined on $\mathcal{D}$ that are symmetric with respect to the two midlines of the rectangle, i.e.,

$$q(\pi/a - x, y) = q(x, y) = q(x, \pi - y).$$

Our spectral data $\boldsymbol{\Lambda}$ are the first N eigenvalues of (1) with the Dirichlet condition $u = 0$ on $\partial\mathcal{D}$. Such an inverse problem was previously considered in [10], where an algorithm based on the Rayleigh–Ritz method was used to construct approximations to a truncated Fourier series expansion of $q(x, y)$.

A piecewise constant approximation to $q(x, y)$ is constructed by dividing $\mathcal{D}$ into a uniform grid of $4M$ subrectangles. D_k will denote a subset of $\mathcal{D}$ consisting of four of these smaller rectangles symmetrically chosen around the midlines $x = \pi/2a$ and

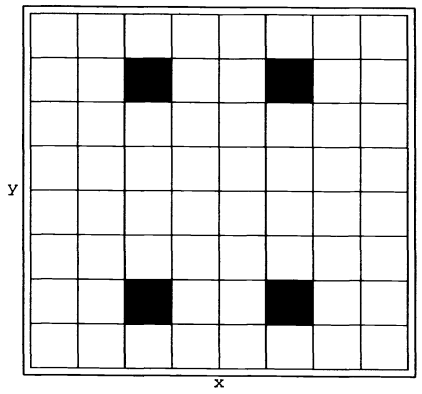


FIG. 1.

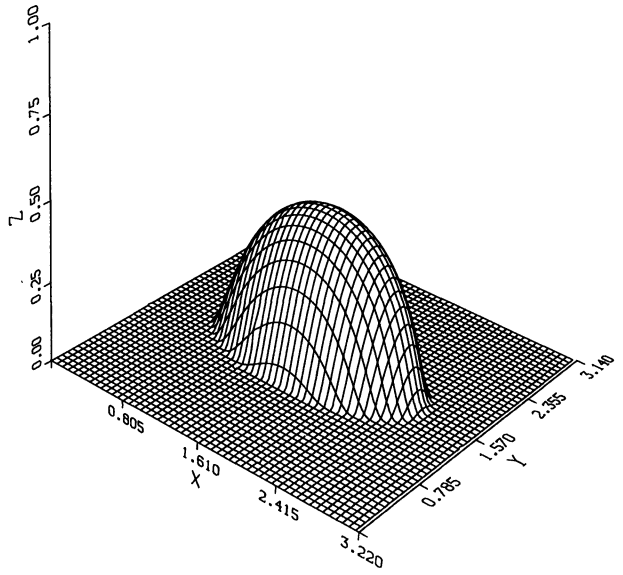


FIG. 2.

$y = \pi/2$ of $\mathcal{D}$ (see Fig. 1). Each of the resulting M step functions $h_k(x, y)$ is then symmetric with respect to the midlines of $\mathcal{D}$.

The rectangle here will be $\mathcal{D} = [0, \pi/\sqrt{0.95}] \times [0, \pi]$. The choice of $a = \sqrt{0.95}$ guarantees that the lowest 50 eigenvalues of $\beta = 0$ are simple, allowing the linearization formula to be used at $\beta_0 = 0$. The function for which we will construct an approximation is as follows:

$$q(x,y) = \begin{cases} \exp(-\frac{1}{d(x,y)}) & \text{if } d(x,y) = (\frac{3\pi}{8})^2 - (x - \frac{\pi}{2\sqrt{0.95}})^2 - 4(y - \frac{\pi}{2}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The graph of this function is shown in Fig. 2. Starting with $\beta_0 = 0$ as an initial guess of a solution to the nonlinear problem (17), up to five iterations of the linear least squares problem (18) were solved over the set $\mathcal{B}(1)$ for $N = 9, 16$, and 25. The subrectangles of $\mathcal{D}$ were formed by dividing $\mathcal{D}$ into a uniform grid of $2\sqrt{N} \times 2\sqrt{N}$

TABLE 1

$N = M$	Grid	$\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q_{N,M})\|_2$	$\|q - q_{N,M}\|_{1Max}$
9	6x6	0.0029	0.180
16	8x8	0.0011	0.050
25	10x10	0.0060	0.023

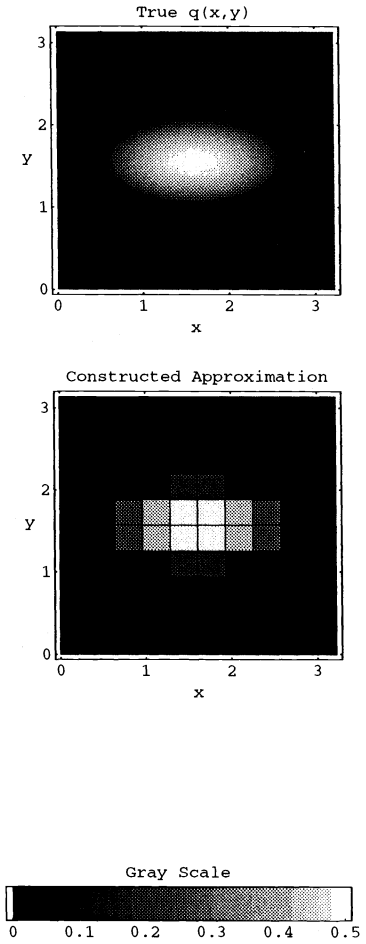


FIG. 3.

for a total of $4M$ rectangles. For each N, the final computed $\boldsymbol{\beta}_i \in \boldsymbol{R}^N$ was used to construct $q_{N,M}$ given by (14). Estimates of the resulting errors $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q_{N,M})\|_2$ and $\|q - q_{N,M}\|_{1Max}$ are shown in Table 1.

Fig. 3 displays a density plot representation of the actual $q(x, y)$ along with a density plot of the piecewise constant approximation constructed on a 10×10 grid using $N = 25$ eigenvalues.

5.3. Recovery of a general potential on a rectangle. Our next example illustrates a generalization of the two-spectra inverse problem for the Sturm–Liouville equation. We will continue to consider $q(x, y)$ defined on a rectangular domain $\mathcal{D}$;

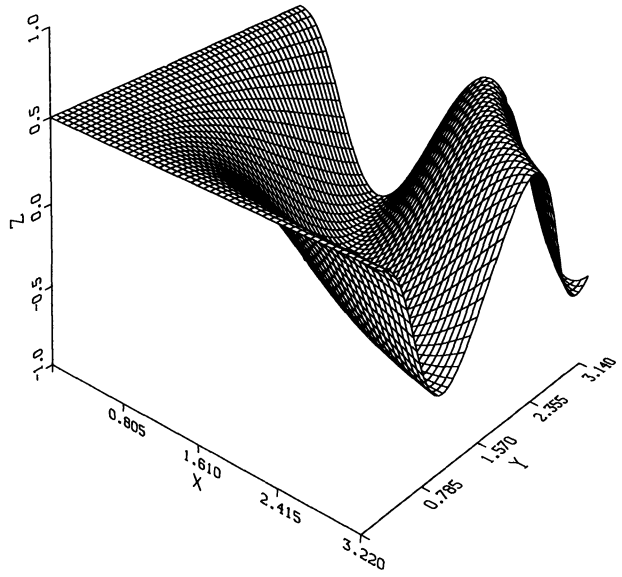


FIG. 4.

however, here $q(x, y)$ is no longer assumed to be symmetric with respect to the midlines of the rectangle. The spectral data $\boldsymbol{\Lambda}$ used to construct an approximation to $q(x, y)$ will consist of eigenvalues from four different sets of boundary conditions. In particular, we use $\boldsymbol{\Lambda} = \cup_{i=1}^{4} \boldsymbol{\Lambda}_i$, where $\boldsymbol{\Lambda}_i$ are eigenvalues resulting from the following boundary conditions:

$$\begin{aligned}
\boldsymbol{\Lambda}_1 &: \quad u(x,0) = u(y,0) = u(x,\pi) = u(\pi/a, y) = 0, \\
\boldsymbol{\Lambda}_2 &: \quad u(x,0) = u(y,0) = u(x,\pi) = u_x(\pi/a, y) = 0, \\
\boldsymbol{\Lambda}_3 &: \quad u(x,0) = u(y,0) = u_y(x,\pi) = u(\pi/a, y) = 0, \\
\boldsymbol{\Lambda}_4 &: \quad u(x,0) = u(y,0) = u_y(x,\pi) = u_x(\pi/a, y) = 0.
\end{aligned}$$

A piecewise constant approximation to $q(x, y)$ is constructed by dividing $\mathcal{D}$ into a uniform grid of M subrectangles D_k. As before, the M basis functions $h_k(x, y)$ are taken to be 1 on D_k and 0 otherwise.

On the rectangle $\mathcal{D} = [0, \pi/\sqrt{0.95}] \times [0, \pi]$, we constructed an approximation to the function

$$q(x, y) = 0.5 \cos(xy).$$

The graph of $q(x, y)$ is shown in Fig. 4. The constructed approximation used $N = 100$ eigenvalues (25 eigenvalues from each boundary condition) and $M = 100$ basis functions (a 10×10 grid of $\mathcal{D}$). Beginning with an initial guess of $\boldsymbol{\beta}_0 = 0$, seven iterations of the linear least squares problem (18) were solved to find $\boldsymbol{\beta}_7 \in \boldsymbol{R}^{100}$. The resulting piecewise constant function $q_{100,100}(x, y)$ was estimated to satisfy

$$\|q - q_{100,100}\|_{1\text{Max}} = 0.033 \quad \|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q_{100,100})\|_2 = 0.0052.$$

Density plots of $q(x, y)$ and $q_{100,100}(x, y)$ are shown in Fig. 5.

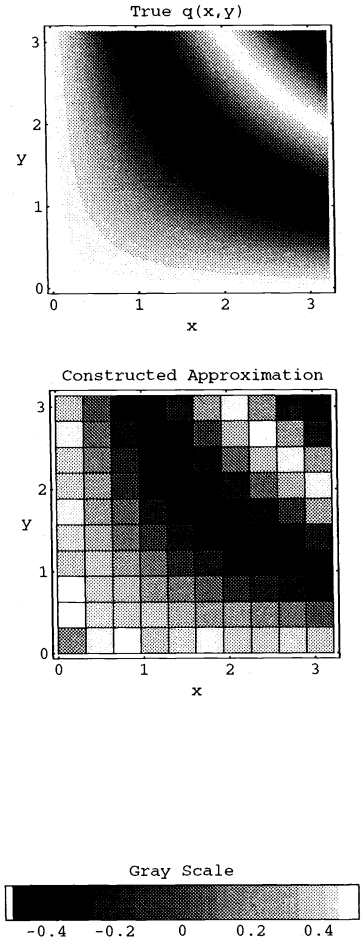


FIG. 5.

5.4. Recovery of a characteristic function. Our final example is a special case of the preceding problems, namely, the reconstruction of a function of the form

$$q(x,y) = \begin{cases} 1 & \text{if } (x,y) \in S, \\ 0 & \text{otherwise,} \end{cases}$$

where S is a subset of $\mathcal{D}$. The algorithm proceeds as in §5.3; however, once $\boldsymbol{\beta}$ has been computed, we then take

$$S_{N,M} = \bigcup_{\beta_k \geq 0.5} D_k$$

as an approximation to S.

Returning to the rectangle $\mathcal{D} = [0, \pi/\sqrt{0.95}] \times [0, \pi]$, we exhibit an approximation to the characteristic function for the region S shown in Fig. 6. This set is comprised of two parts, namely, the lower half of a disk with radius 0.5 centered at the center of $\mathcal{D}$, and a square with side length 0.2 centered at $(3\pi/4\sqrt{0.95}, 3\pi/4)$.

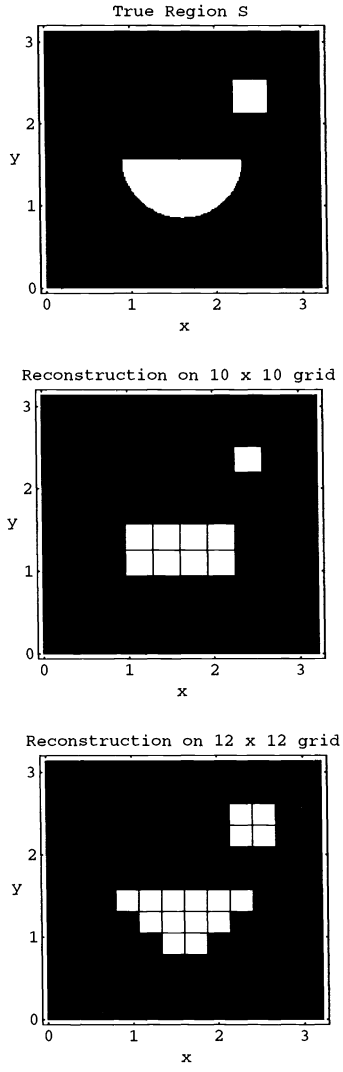


FIG. 6.

Using 100 eigenvalues (25 from each boundary condition), two approximations to S were constructed, one on a 10×10 mesh and a second on a 12×12 mesh. The resulting approximations $S_{100,100}$ and $S_{100,144}$ are shown in Fig. 6.

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] M. S. ASHBAUGH, E. M. HARRELL II, AND R. SVIRSKY, *On minimal and maximal eigenvalue gaps and their causes*, Pacific J. Math., 147 (1991), pp. 1–24.
[3] V. BARCILON, *A two-dimensional inverse eigenvalue problem*, Inverse Problems, 6 (1990), pp. 11–20.

[4] D. C. BARNES, *The inverse eigenvalue problem with finite data*, SIAM J. Math. Anal., 22 (1991), pp. 732–753.
[5] V. DUBROVSKII AND A. NAGORNYI *The inverse problem for the degree of the laplace operator with potential in* L^2, Differentsial'nye Uravneniya, 28 (1992), pp. 1552–1561.
[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Interscience, New York, 1958.
[7] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, San Francisco, CA, 1969.
[8] R. J. HANSON AND K. H. HASKEL, *ALGORITHM 587. Two algorithms for the linearly constrained least squares problem*, ACM Trans. Math. Software, 8 (1982), pp. 323–333.
[9] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1976.
[10] R. KNOBEL AND J. R. MCLAUGHLIN, *A reconstruction method for a two dimensional inverse eigenvalue problems*, preprint.
[11] J. R. MCLAUGHLIN, *Analytical methods for recovering coefficients in differential equations from spectral data*, SIAM Rev., 28 (1986), pp. 53–72.
[12] A. NACHMAN, J. SYLVESTER, AND G. UHLMANN, *An n-dimensional Borg-Levinson theorem*, Comm. Math. Phys., 115 (1988), pp. 595–605.
[13] T. SEIDMAN, *An inverse eigenvalue problem with rotational symmetry*, Inverse Problems, 4 (1988), pp. 1093–1115.
[14] W. STENGER, *On the variational principles for eigenvalues for a class of unbounded operators*, J. Math. Mech., 17 (1968), pp. 641–648.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 633–658, May 1995

007

FREE BOUNDARY PROBLEMS FOR POTENTIAL AND STOKES FLOWS VIA NONSMOOTH ANALYSIS*

SRDJAN STOJANOVIC† AND THOMAS SVOBODNY‡

Abstract. A new approach to some free boundary problems of the type of jets and cavities for potential flows is introduced. Both potential *and* Stokes flows are considered. The variable domain problems are relaxed so that they become nonsmooth optimization problems on fixed domains for somewhat singular state equations. State equations are considered, and multivalued generalized gradients of the variational functionals are studied. The method is constructive.

Key words. free boundary, Stokes problem, nonsmooth analysis

AMS subject classification. 35Q

1. Introduction. Consider the following now-classical variational problem introduced and solved by Alt and Caffarelli [2], and studied extensively by Alt, Caffarelli, and Friedman. (See [3] and [8] and references given there.) (See also [14] for numerical considerations; for the simplicity of presentation we discuss the very particular geometry: $\Omega = (-a, a) \times (0, 2)$.)

Find $w \in H^1(\Omega)$ satisfying the boundary conditions

$$w = 0 \text{ in } \{(x,0); -a < x < a\},\ w = 2 \text{ in } \{(x,2); -a < x < a\} \tag{1.1.1}$$

such that the variational functional

$$\mathbf{J}(w) = \int_\Omega \left[|\nabla w|^2 + g^2 \mathbf{I}_{\{w>0\}}\right] \tag{1.1.2}$$

is minimized. Here $\mathbf{I}_D$ is a characteristic function of the set D, i.e.,

$$\mathbf{I}_D(x) = \begin{cases} 1 & \text{if } x \in D, \\ 0 & \text{if } x \notin D, \end{cases} \tag{1.1.3}$$

and $g \geq 0$ is a given function.

It is well known (see [2]) that, under certain conditions, a minimizer w satisfies

$$\begin{gathered} \Delta w = 0 \text{ in } \Omega \cap \{w > 0\}, \\ |\nabla w| = g,\ w = 0 \text{ in } \Omega \cap \partial\{w > 0\}, \\ w = 0 \text{ in } \{(x,0); -a < x < a\},\ w = 2 \text{ in } \{(x,2); -a < x < a\}, \\ w_x = 0 \text{ in } \{(\pm a, y); 0 < y < 2\}. \end{gathered} \tag{1.1.4}$$

REMARK 1.1.1. *Moreover, if $g_y \leq 0$ then, using monotone rearrangemets (see [11]) one can easily show that that there exists a minimizer w such that $w_y \geq 0$. That implies that there exists a function $u = u(x)$ such that $\Omega \cap \partial\{w > 0\} = \{(x, u(x)); -a < x < a\}$. Furthermore, if $g \in C^{k,\alpha}(\Omega)$ then, by the theorem of Alt and Caffarelli, $u \in C^{k+1,\alpha}$.*

* Received by the editors April 2, 1993; accepted for publication (in revised form) November 15, 1993.

† Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221-0025. This author was supported in part by National Science Foundation grant DMS-91-11794.

‡ Department of Mathematics and Statistics, Wright-State University, Dayton, Ohio 45435. This author was supported in part by Office of Naval Research grant N00014-91-1494.

The applications of this problem are mainly in potential fluid mechanics, i.e., in free boundary problems of the type of jets and cavities for potential flows. (See [8] and the references given there.)

Unfortunately, the above variational approach failed, in the case of Stokes and Navier–Stokes equation. The reason is that there is no known analog of the functional (1.1.2).

In this paper we introduce a new approach to this problem. The approach is discussed in the case of potential flow (§2), and in the case of the Stokes flow (§3). Results of the §2 were announced by Stojanovic in [15]. See [4] for a related method.

2. Potential flow.

2.1. Statement of the problem. To further motivate our approach, we observe that the "Euler equation" for the minimizer of the functional (1.1.2) is

$$\begin{aligned} \Delta w &= \xi_{\Omega\cap\partial\{w>0\}} \text{ in } \mathcal{D}'(\Omega), \\ w = 0 \text{ in } \{(x,0); -a < x < a\},\ & w = 2 \text{ in } \{(x,2); -a < x < a\}, \\ w_x &= 0 \text{ in } \{(\pm a, y); 0 < y < 2\}, \end{aligned} \tag{2.1.1}$$

where for any regular surface Γ, the measure ξ_Γ is defined by

$$\xi_\Gamma(\varphi) = \int_\Gamma g\varphi d\sigma. \tag{2.1.2}$$

Of course, (2.1.1) is a very difficult equation since the measure on the right-hand side depends on a solution. On the other hand, if the right-hand side does not depend on a solution, i.e., if the equation is merely

$$\begin{aligned} \Delta w &= \xi_\Gamma \text{ in } \mathcal{D}'(\Omega), \\ w = 0 \text{ in } \{(x,0); -a < x < a\},\ & w = 2 \text{ in } \{(x,2); -a < x < a\}, \\ w_x &= 0 \text{ in } \{(\pm a, y); 0 < y < 2\}, \end{aligned} \tag{2.1.3}$$

for some given (fixed) regular surface Γ, then (2.1.3) is a fairly simple equation. So the idea is to study (2.1.3) and then to look for Γ such that if w is the corresponding solution of (2.1.3), then

$$\Gamma = \Omega \cap \partial\{w > 0\}. \tag{2.1.4}$$

So, consider the set of admissible shapes (see Remark 1.1.1)

$$U = \left\{u \in H_0^3(-1,1); 0 \le u(x) \le 1, -1 < x < 1\right\}. \tag{2.1.5}$$

Denote

$$\Gamma_u = \{(x, u(x)); -1 < x < 1\}, \tag{2.1.6}$$

and extend $u \in U$ as zero outside of $(-1,1)$. Define the domain

$$\Omega_u = \{(x,y);\ |x| < a,\ u(x) < y < 2\}. \tag{2.1.7}$$

Let $w = w^u$, be the solution of

$$\begin{aligned} \Delta w &= 0 \text{ in } \Omega_u, \\ w = 0 \text{ in } \{(x,0); -a < x < -1 \text{ or } 1 < x < a\},\ & w = 0 \text{ in } \Gamma_u, \\ w = 2 \text{ in } \{(x,2); -a < x < a\},\ & w_x = 0 \text{ in } \{(\pm a, y); 0 < y < 1\}. \end{aligned} \tag{2.1.8}$$

We could also take $a = \infty$ in (2.1.8), i.e., consider a flow in an infinite channel. Then the last condition in (2.1.8) is substituted by the requirement that w is bounded.

In the context of potential fluid mechanics, w is a stream function. If a stream function w is known then, of course, the velocity vector field $\mathbf{v}$ can be computed easily as $\mathbf{v} = \langle w_y, -w_x \rangle$.

The problem we propose is the following.

For given $g = g(x, y)$ such that (we will not always have to assume this much)

$$g \in C^{1,1}(\Omega), \tag{2.1.9}$$

$$g = 0 \text{ in } \Omega \cap \{|x| > 1\}, \tag{2.1.10}$$

find (if possible) $u \in U$ such that, if w^u is the corresponding solution of (2.1.8), then also

$$|\mathbf{v}| = |\nabla w^u| = g \text{ in } \Gamma_u. \tag{2.1.11}$$

We note that by the Bernoulli's law

$$P + \frac{1}{2}|\nabla w^u|^2 = \text{const} \tag{2.1.12}$$

throughout the fluid (here P denotes the pressure). Hence, we see that requesting specific velocity profile on the immersed obstacle is equivalent to requesting the specific pressure (and hence, force) profile. In §3 we shall study the exactly analogous problem for the Stokes equation. So the method introduced here, although not as satisfactory as the variational method of Alt and Caffarelli [2], is applicable to more important equations.

2.2. A relaxation of the problem. Suppose that there exists an $u \in U$ such that corresponding w^u solves (2.1.8) and (2.1.11). We shall say then that u is an *exact shape*. Now, extend w^u from Ω_u to Ω as z^u:

$$z^u = \begin{cases} w^u & \text{on } \Omega_u, \\ 0 & \text{on } \Omega \backslash \Omega_u. \end{cases} \tag{2.2.1}$$

Lemma 2.2.1 follows.

LEMMA 2.2.1. *If $u \in U$ is an exact shape, then $z^u \in H^1(\Omega)$, and it is a solution of the following elliptic boundary value problem (with singular right-hand side):*

$$\begin{gathered} \Delta z^u = \xi_u \text{ in } \Omega, \\ z^u = 0 \text{ in } \{(x,0); -a < x < a\},\ z^u = 2 \text{ in } \{(x,2); -a < x < a\}, \\ (z^u)_x = 0 \text{ in } \{(\pm a, y); 0 < y < 1\}, \end{gathered} \tag{2.2.2}$$

where $\xi_u \in H^{-1}(\Omega)$ is a measure given by

$$\xi_u(\varphi) = \int_{\Gamma_u} g\varphi d\sigma. \tag{2.2.3}$$

Proof. The proof is obvious, since by elliptic estimates w^u is regular in Ω_u, $z^u \in C^{0,1}(\bar{\Omega})$ (regarding regularity near corners, see the beginning of the proof of the Theorem 2.3.1), and in particular $z^u \in H^1(\Omega)$.

By the trace theorem,

$$|\xi_u(\varphi)| \le \|g\|_{L^2(\Gamma_u)}\|\varphi\|_{L^2(\Gamma_u)} \le c_u\|g\|_{H^1(\Omega)}\|\varphi\|_{H^1(\Omega)}. \tag{2.2.4}$$

So, in particular, $\xi_u \in H^{-1}(\Omega)$. Also, since $g \ge 0$, ξ_u is a measure.

Now, more explicitly, (2.2.2) can be written as the following: Find $z \in H^1(\Omega)$ such that

$$z^u = 0 \text{ in } \{(x,0); -a < x < a\},\ z^u = 2 \text{ in } \{(x,2); -a < x < a\} \tag{2.2.5}$$

and

$$-\int_\Omega \nabla z^u \cdot \nabla\varphi = \int_{\Gamma_u} g\varphi d\sigma \tag{2.2.6}$$

for all $\varphi \in H^1(\Omega)$ such that

$$\varphi = 0 \text{ in } \{(x,0); -a < x < a\} \cup \{(x,2); -a < x < a\}. \tag{2.2.7}$$

To check (2.2.6), we note that by the maximum principle, a solution of (2.1.8) is positive. Hence (2.1.11) and the boundary condition in (2.1.8) imply that

$$\frac{\partial w^u}{\partial \nu_u} = -g \text{ in } \Gamma_u, \tag{2.2.8}$$

where ν_u is the exterior unit normal to $\partial\Omega_u$. Hence,

$$\begin{aligned} -\int_\Omega \nabla z^u \cdot \nabla\varphi &= -\int_{\Omega_u} \nabla w^u \cdot \nabla\varphi \\ &= \int_{\Omega_u} (\Delta w^u)\varphi - \int_{\partial\Omega_u} \frac{\partial w^u}{\partial \nu_u}\varphi d\sigma = \int_{\Gamma_u} g\varphi d\sigma, \end{aligned} \tag{2.2.9}$$

which completes the proof of the lemma.

LEMMA 2.2.2. *Let z^u be a solution of* (2.2.5)–(2.2.7). *If it happens that $z^u|_{\Gamma_u} = 0$, then $z^u|_{\Omega_u}$ is a solution of* (2.1.8)–(2.1.11), *i.e., u is an exact shape.*

Proof. In the next section we shall prove that z^u is regular enough so that calculations performed here are legitimate. More precisely, by (2.3.6) below, it suffices to assume that $\varphi \in C_0^1(\Omega)$. We have

$$\begin{aligned} \int_{\Gamma_u} g\varphi d\sigma = &-\int_{\Omega_u} \nabla z^u \cdot \nabla\varphi - \int_{\Omega\setminus\Omega_u} \nabla z^u \cdot \nabla\varphi = \int_{\Omega_u} (\Delta z^u)\varphi \\ &+ \int_{\Omega\setminus\Omega_u} (\Delta z^u)\varphi - \int_{\partial\Omega_u} \frac{\partial z^u}{\partial \nu}\varphi d\sigma - \int_{\partial(\Omega\setminus\Omega_u)} \frac{\partial z^u}{\partial \nu}\varphi d\sigma. \end{aligned} \tag{2.2.10}$$

Let ν be exterior to Ω_u, and let

$$z^{u,\text{int}} \stackrel{\text{def}}{=} z^u|_{\Omega\setminus\Omega_u},\ z^{u,\text{ext}} \stackrel{\text{def}}{=} z^u|_{\Omega_u}. \tag{2.2.11}$$

Then (2.2.10) implies that

$$\int_{\Gamma_u} g\varphi d\sigma = \int_{\Gamma_u} \left(\frac{\partial z^{u,\text{int}}}{\partial \nu} - \frac{\partial z^{u,\text{ext}}}{\partial \nu}\right)\varphi d\sigma,\ \forall \varphi \in C_0^1(\Omega). \tag{2.2.12}$$

So,

$$g = \frac{\partial z^{u,\text{int}}}{\partial \nu} - \frac{\partial z^{u,\text{ext}}}{\partial \nu} \text{ on } \Gamma_u. \tag{2.2.13}$$

We observe that (2.2.13) always holds for the solution of (2.2.2).

Now, if $z^u|_{\Gamma_u} = 0$, then $z^u|_{\Omega \setminus \Omega_u} = 0$, so that $(\partial z^{u,\text{int}})/\partial \nu = 0$, and then $g = -(\partial z^{u,\text{ext}})/\partial \nu$ on Γ_u, i.e.,

$$g = |\nabla (z^u|_{\Omega_u})| \text{ on } \Gamma_u, \tag{2.2.14}$$

i.e., (2.1.11) holds.

Lemma 2.2.2 motivates the following.

DEFINITION 2.2.1. *$u^* \in U$ is said to solve the* relaxed free boundary problem *if the corresponding z^u defined by* (2.2.2) *is such that*

$$\Phi(u) = \frac{1}{2} \int_{\Gamma_u} (z^u)^2 d\sigma \tag{2.2.15}$$

is minimized, i.e., $u^* \in U$ is such that

$$\Phi(u^*) = \min_{u \in U} \Phi(u). \tag{2.2.16}$$

Of course, an exact shape is a minimizer, i.e., a solution of (2.2.16). On the other hand, a solution of (2.2.16) is an exact shape, provided an exact shape exists.

We do not consider whether an exact shape exists. Rather, we shall study the relaxed problem introduced in Definition 2.2.1.

2.3. The state equation. It will be convenient to state the regularity theorem for the general boundary value. So let ψ be a given function on Ω such that $z^u = \psi$ on $\partial_0 \Omega \subset \partial \Omega$. We assume that the boundary and ψ are sufficiently regular (see [16] for details; also, we shall give some details in the case of the boundary and boundary values in our case). For any $z \in H^1(\Omega)$, we define $\|z\|_{L^\infty(\partial\Omega)}$ as

$$\|z\|_{L^\infty(\partial\Omega)} \overset{\text{def}}{=} \inf \{m \geq 0; -m \leq z \leq m \text{ on } \partial\Omega \text{ in } H^1(\Omega)\}, \tag{2.3.1}$$

where inequalities in $H^1(\Omega)$ are defined in, e.g., [16]. Also, we define

$$W^{2,q}_{\{y>0\}-loc}(\Omega) \overset{\text{def}}{=} \cap_{\epsilon>0} W^{2,q}(\Omega \cap \{y > \epsilon\}). \tag{2.3.2}$$

We have the next theorem.

THEOREM 2.3.1. *For any $u \in U$ the state equation* (2.2.2) *has a unique weak solution. Let q be such that $2 \leq q < \infty$. If $g \in W^{1,q}(\Omega)$, then*

$$z^u \in W^{1,q}(\Omega) \cap C^\infty(\bar{\Omega} \setminus \Gamma_u), \tag{2.3.3}$$

and the a priori estimate

$$\|z^u\|_{W^{1,q}(\Omega)} \leq c\left(1 + \|u\|_{C^{0,1}(-1,1)}\right)\left(\|g\|_{W^{1,q}(\Omega)} + \|\psi\|_{W^{1,q}(\Omega)}\right). \tag{2.3.4}$$

holds. If in addition $q > 2$, then

$$\|z^u\|_{L^\infty(\Omega)} \leq c\left(1 + \|u\|_{C^{0,1}(-1,1)}\right)\left(\|g\|_{W^{1,q}(\Omega)} + \|\psi\|_{L^\infty(\partial\Omega)}\right). \tag{2.3.5}$$

Moreover, if $g \in W^{2,q}(\Omega)$, *and* (2.1.10) *holds, then* (*see* (2.2.11))

$$z^{u,\mathrm{ext}} \in W^{2,q}(\Omega_u),\ z^{u,\mathrm{int}} \in W^{2,q}_{\{y>0\}-\mathrm{loc}}(\Omega \setminus \bar{\Omega}_u) \tag{2.3.6}$$

and the a priori estimates

$$\|z^{u,\mathrm{ext}}\|_{W^{2,q}(\Omega_u)} \le c\left(\|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right) \tag{2.3.7}$$

and

$$\begin{aligned} &\|z^{u,\mathrm{int}}\|_{W^{2,q}((\Omega\setminus\Omega_u)\cap\{y>\epsilon\})} \\ &\le c\left(\epsilon, \|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right) \end{aligned} \tag{2.3.8}$$

hold.

Proof. Since $\xi_u \in H^{-1}(\Omega)$ existence and uniqueness of a weak solution z^u of (2.2.2) is trivial. Also, since z^u is harmonic in $\Omega\setminus\Gamma_u$, it follows that $z^u \in C^\infty(\bar{\Omega}\setminus\Gamma_u)$. Few words are needed here because of the presence of corners in Ω. To prove regularity of z^u in the neighborhood of corners, say, in the neighborhood of $(-a,0)$, one can extend z^u in $\{x<-a, 0<y<2\}$ as $\widetilde{z^u}$ by the formula

$$\widetilde{z^u}(x,y) \overset{\mathrm{def}}{=} \begin{cases} z^u(-2a-x,y) & \text{if } x<-a, \\ z^u(x,y) & \text{if } x \ge -a. \end{cases} \tag{2.3.9}$$

Then since $\widetilde{z^u}$ is continuous on $\{x=-a\}$ and $\widetilde{z^u}_x = 0$ on $\{x=-a\}$, it is elementary to show that $\widetilde{z^u}$ is harmonic across $\{x=-a\}$. Indeed, let $B_\rho(A) = B_1 \cup (B_\rho(A)\cap\{x=-a\}) \cup B_2 \subset \{0<y<2\}$ be a ball centered at $A \in \{x=-a\}$ with radius ρ. Here, $B_1 = B_\rho(A)\cap\{x>-a\}$ and $B_2 = B_\rho(A)\cap\{x<-a\}$. Then,

$$\begin{aligned} \int_{B_\rho(A)} \widetilde{z^u}\Delta\varphi &= \int_{B_1} \widetilde{z^u}\Delta\varphi + \int_{B_2} \widetilde{z^u}\Delta\varphi \\ &= \int_{\{x=-a\}\cap B_\rho(A)} \left[-\varphi_x \widetilde{z^u} + \varphi \widetilde{z^u}_x + \varphi_x \widetilde{z^u} - \varphi \widetilde{z^u}_x\right] dy = 0 \end{aligned} \tag{2.3.10}$$

for all $\varphi \in C_0^\infty(B_\rho(A))$, so that $\widetilde{z^u}$ is harmonic across $\{x=-a\}$ as claimed. Henceforth $\widetilde{z^u}$ is as regular in the neighborhood of $(-a,0)$ as the (extended) boundary data is. In our case the boundary data is $\psi = 0$, so that (2.3.3) follows.

Set $\varphi = \psi - z^u$ in (2.2.6). It easily follows that

$$\begin{aligned} \int_\Omega |\nabla z^u|^2 &= \int_{\Gamma_u} g(\psi - z^u)d\sigma - \int_\Omega \nabla z^u \cdot \nabla\psi \\ &\le \left(\int_{\Gamma_u} g^2 d\sigma\right)^{\frac{1}{2}} \left[\left(\int_{\Gamma_u} \psi^2 d\sigma\right)^{\frac{1}{2}} + \left(\int_{\Gamma_u} (z^u)^2 d\sigma\right)^{\frac{1}{2}}\right] \\ &\qquad + \left|\int_\Omega \nabla z^u \cdot \nabla\psi\right|. \end{aligned} \tag{2.3.11}$$

Now since $z^u = (z^u - \psi) + \psi$, using Poincaré inequality, we have

$$\begin{aligned} \|z^u\|_{H^1(\Omega)} &\le c\left(\|\nabla(z^u-\psi)\|_{L^2(\Omega)} + \|\psi\|_{H^1(\Omega)}\right) \\ &\le c\left(\|\nabla z^u\|_{L^2(\Omega)} + \|\psi\|_{H^1(\Omega)}\right). \end{aligned} \tag{2.3.12}$$

Combining (2.3.11) and (2.3.12), we get

$$\|z^u\|^2_{H^1(\Omega)} \le c\left(1+\|u\|_{C^{0,1}(-1,1)}\right)\left[\|g\|_{H^1(\Omega)}\left(\|\psi\|_{H^1(\Omega)}+\|z^u\|_{H^1(\Omega)}\right)\right]$$
$$+c\|z^u\|_{H^1(\Omega)}\|\psi\|_{H^1(\Omega)}+c\|\psi\|^2_{H^1(\Omega)}. \tag{2.3.13}$$

In (2.3.13) the inequality follows from the proof of the trace theorem (see, e.g., [12] or [7]). Indeed, one can see ([7], p. 132) that for $1 \le q < \infty$, one has

$$\|z^u\|^q_{L^q(\Gamma_u)} \le c\left(1+\|u\|^2_{C^{0,1}(-1,1)}\right)^{\frac{1}{2}}\|z^u\|^q_{W^{1,q}(\Omega)}, \tag{2.3.14}$$

which implies

$$\|z^u\|_{L^q(\Gamma_u)} \le c\left(1+\|u\|_{C^{0,1}(-1,1)}\right)^{\frac{1}{q}}\|z^u\|_{W^{1,q}(\Omega)}. \tag{2.3.15}$$

From (2.3.13) we easily conclude that (2.3.4) holds for $q=2$.

Proceeding, we assume $\frac{1}{q}+\frac{1}{q'}=1$ and

$$\begin{aligned}|\xi_u(\varphi)| &\le \|g\|_{L^q(\Gamma_u)}\|\varphi\|_{L^{q'}(\Gamma_u)}\\ &\le c\left(1+\|u\|_{C^{0,1}(-1,1)}\right)^{\frac{1}{q}}\|g\|_{W^{1,q}(\Omega)}\left(1+\|u\|_{C^{0,1}(-1,1)}\right)^{\frac{1}{q'}}\|\varphi\|_{W^{1,q'}(\Omega)}\\ &= c\left(1+\|u\|_{C^{0,1}(-1,1)}\right)\|g\|_{W^{1,q}(\Omega)}\|\varphi\|_{W^{1,q'}(\Omega)}. \end{aligned}\tag{2.3.16}$$

So, $\xi_u \in (W^{1,q'}(\Omega))^*$ (here X^* represents the dual space of the space X) and

$$\|\xi_u\|_{(W^{1,q'}(\Omega))^*} \le c\left(1+\|u\|_{C^{0,1}(-1,1)}\right)\|g\|_{W^{1,q}(\Omega)}. \tag{2.3.17}$$

We know (see, e.g., [1]) that ξ has a representation $\xi(\varphi)=\int_\Omega [f_0\varphi+f_1\varphi_x+f_2\varphi_y]$, for some $f_i \in L^q(\Omega), i=0,1,2$, and

$$\|\xi_u\|_{(W^{1,q'}(\Omega))^*} = \sum_{i=0}^{2}\|f_i\|_{L^q(\Omega)}. \tag{2.3.18}$$

Now from elliptic regularity (see [16], p. 179), we have

$$\|z^u\|_{W^{1,q}(\Omega)} \le c\left(\sum_{i=0}^{2}\|f_i\|_{L^q(\Omega)}+\|\psi\|_{W^{1,q}(\Omega)}+\|z^u\|_{H^1(\Omega)}\right). \tag{2.3.19}$$

From (2.3.17)–(2.3.19) and since (2.3.4) is already proved in the case $q=2$, we conclude that (2.3.4) holds.

To prove (2.3.5), we recall (see, e.g., [16], p. 103) that if $q>2$ and if $z^u \le 0$ on $\partial_0\Omega$ in the sense of $H^1(\Omega)$, then

$$\operatorname*{ess\,sup}_{\Omega} z^u \le c\left(\sum_{i=0}^{2}\|f_i\|_{L^q(\Omega)}+\|z^u\|_{L^2(\Omega)}\right). \tag{2.3.20}$$

Hence,

$$\begin{aligned}&\operatorname*{ess\,sup}_{\Omega}\left(z^u-\|z^u\|_{L^\infty(\partial\Omega)}\right)\\ &\le c\left(\sum_{i=0}^{2}\|f_i\|_{L^q(\Omega)}+\|z^u\|_{L^2(\Omega)}+\|z^u\|_{L^\infty(\partial\Omega)}\right),\end{aligned}\tag{2.3.21}$$

and similarly for $-z^u + \|z^u\|_{L^\infty(\partial\Omega)}$. This easily implies (2.3.5).

Now, we shall consider further regularity of $z^u|_{\Omega_u}$ and $z^u|_{\Omega\setminus\Omega_u}$. Since the singular set is on Γ_u, we expect higher regularity in the tangential direction. To prove that this is the case we flatten the Γ_u first, since then it is easier to differentiate.

Define v, $\tilde{g}$, and $\tilde{\varphi}$ by $v(x,y) = z^u(x, y+u(x))$, $\tilde{g}(x,y) = g(x, y+u(x))\sqrt{1+u'^2(x)}$, $\tilde{\varphi}(x,y) = \varphi(x, y+u(x))$ and operator L by $Lv = \Delta v + v_{yy}(u_x)^2 - 2v_{xy}u_x - v_y u_{xx}$. Of course, L is uniformly elliptic, since the matrix

$$[l_{ij}] = \begin{bmatrix} 1 & -u_x \\ -u_x & 1+u_x^2 \end{bmatrix} \tag{2.3.22}$$

is positive definite. Indeed, $l_{ij}\xi_i\xi_j = (\xi_1 - u_x\xi_2)^2 + \xi_2^2$. So, if c is such that $|u_x| \le c$, then if $|\xi_2| < \frac{1}{2c}|\xi_1|$ then $(\xi_1 - u_x\xi_2)^2 > \frac{1}{4}\xi_1^2$. On the other hand, if $|\xi_2| \ge \frac{1}{2c}|\xi_1|$ then $\xi_2^2 \ge \frac{1}{4c^2}\xi_1^2$. So, it is easy to see that if we take $\alpha = \min\left(\frac{1}{4}, \frac{1}{8c^2}\right)$ then $l_{ij}\xi_i\xi_j \ge \alpha|\xi|^2$.

Also let Ξ_u be the map with the image Ω given by the formula

$$\Xi_u(x,y) = (x, y+u(x)). \tag{2.3.23}$$

Then, $\Delta z^u \circ \Xi_u = Lv$, and since $|\det D\Xi_u| = 1$ (here $D\Xi_u$ is the gradient matrix of the map Ξ_u so that $|\det D\Xi_u|$ is the Jacobian)

$$(Lv)(\tilde{\varphi}) = (\Delta z^u)(\varphi). \tag{2.3.24}$$

Hence

$$\begin{aligned} (Lv)(\tilde{\varphi}) &= \int_{\Gamma_u} g\varphi d\sigma = \int g(x,u(x))\varphi(x,u(x))\sqrt{1+u'^2(x)}dx \\ &= \int_{\{y=0\}} \tilde{g}\tilde{\varphi}dx \stackrel{\text{def}}{=} \tilde{\xi}(\tilde{\varphi}). \end{aligned} \tag{2.3.25}$$

So

$$Lv = \tilde{\xi} \tag{2.3.26}$$

in the sense of distributions. Since the singular set is now on $\{y=0\}$, we expect higher regularity in x direction. To prove that, we want to differentiate (or more precisely, difference) equation (2.3.26) with respect to x. Somewhat more precisely, define the standard difference operator (in the x direction) δ_h^1 as

$$\left(\delta_h^1 u\right)(x) = \frac{1}{h}\left(u(x+h,y) - u(x,y)\right), \; h \neq 0. \tag{2.3.27}$$

Then from (2.3.26) we get

$$(Lv)\left(\delta_{-h}^1\tilde{\varphi}\right) = \tilde{\xi}\left(\delta_{-h}^1\tilde{\varphi}\right). \tag{2.3.28}$$

We shall discuss in some details only the right-hand side. We have

$$\begin{aligned} \tilde{\xi}\left(\delta_{-h}^1\tilde{\varphi}\right) &= \int_{\{y=0\}} \tilde{g}\delta_{-h}^1\tilde{\varphi}dx \\ &= -\int_{\{y=0\}} \left(\delta_h^1\tilde{g}\right)\tilde{\varphi}dx \longrightarrow -\int_{\{y=0\}} \tilde{g}_x\tilde{\varphi}dx, \end{aligned} \tag{2.3.29}$$

as $h \to 0$. We conclude that $(Lv)_x(\tilde{\varphi}) = \tilde{\xi}_x(\tilde{\varphi})$, and hence

$$L_1 v_x = \tilde{\xi}_x - v_{yy} 2u_x u_{xx} + v_y u_{xxx}, \tag{2.3.30}$$

where $L_1 w = \Delta w + (u_x)^2 w_{yy} - 2u_x w_{xy} - 3u_{xx} w_y$, and where $\tilde{\xi}_x(\tilde{\varphi}) \stackrel{\text{def}}{=} \int_{\{y=0\}} \tilde{g}_x \tilde{\varphi} dx$. We observe that the differencing performed above is legitimate, since

$$\tilde{\xi}_x - v_{yy} 2u_x u_{xx} + v_y u_{xxx} \in \left(W^{1,q'}\right)^*. \tag{2.3.31}$$

Indeed, $g \in W^{2,q}(\Omega)$; also, observe that $u_{xxx} \in L^2$ and that $L^2 \hookrightarrow \left(W^{1,q'}\right)^*$. Also, since L_1 has the same principal part as L, L_1 is uniformly elliptic as well.

Now we can conclude from (2.3.30)–(2.3.31) that $v_x \in W^{1,q}$. This implies, by the trace theorem, that $v_x|_{\{y=0\}} \in W^{1-\frac{1}{q},q}$, so that

$$v|_{\{y=0\}} \in W^{2-\frac{1}{q},q}. \tag{2.3.32}$$

We observe that because of (2.1.5) and (2.1.10), the preceding analysis is true also in the $\{y > 0\}$ neighborhood of (the preimage of) $(\pm 1, 0)$, so that (2.3.32) holds up to the initial and terminal points of (the preimage of) Γ_u. Elliptic regularity then yields $v|_{\{y \geq 0\}} \in W^{2,q}$.

Unfortunately, we cannot claim the same global result for $v|_{\{y \leq 0\}}$ because of the nonsmoothness of $\partial(\Omega \setminus \Omega_u)$, i.e., we have to localize in $\{y < 0\}$. This concludes the proof of (2.3.6). Now, regarding estimates (2.3.7) and (2.3.8), we have

$$\begin{aligned} \|z^{u,\text{ext}}\|_{W^{2,q}(\Omega_u)} &\leq c\left(\|u\|_{H^3(-1,1)}\right) \|v|_{\{y \geq 0\}}\|_{W^{2,q}\left(\Xi_u^{-1}(\Omega_u)\right)} \\ &\leq c\left(\|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right), \end{aligned} \tag{2.3.33}$$

and similarly (after localization in $\{y < 0\}$) for $z^{u,\text{int}}$, which completes the proof of the theorem.

COROLLARY 2.3.1. *If $g \in W^{2,q}(\Omega)$ for some $q > 2$, and if* (2.1.10) *holds, then $z^u \in C^{0,1}_{\{y>0\}-\text{loc}}(\bar{\Omega})$ and the following a priori estimate holds:*

$$\|z^u\|_{C^{0,1}(\bar{\Omega} \cap \{y \geq \epsilon\})} \leq c\left(\epsilon, \|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right), \tag{2.3.34}$$

for any $\epsilon > 0$.

Proof. From (2.3.7) and (2.3.8) and by the imbedding theorem (see, e.g., [9]), we have

$$\begin{aligned} &\|z^{u,\text{ext}}\|_{C^1(\bar{\Omega}_u)} + \|z^{u,\text{int}}\|_{C^1(\overline{\Omega \setminus \Omega_u} \cap \{y \geq 0\})} \\ &\quad \leq c\left(\epsilon, \|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right). \end{aligned} \tag{2.3.35}$$

This implies (2.3.34).

COROLLARY 2.3.2. *If $g \in W^{2,q}(\Omega)$ for some $q > 2$, and if* (2.1.10) *holds, then*

$$z_y^{u,\text{int}} \in C^{0,1-\frac{2}{q}}\left(\overline{\Gamma_u}\right), \tag{2.3.36}$$

and the following a priori estimate

$$\|z_y^{u,\text{int}}\|_{C^{0,1-\frac{2}{q}}\left(\overline{\Gamma_u}\right)} \leq c\left(\|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right) \tag{2.3.37}$$

holds.

The interest in this corollary is due to the lack of $(\Omega \setminus \Omega_u)$-global regularity of z^u.

Proof. Let τ and ν be unit tangent and unit normal to Γ_u. More precisely, set $\tau = 1/(\sqrt{1+u'^2})\,\langle 1, u'\rangle$, and $\nu = 1/(\sqrt{1+u'^2})\,\langle u', -1\rangle$. It is elementary to compute that then $z_y^{u,\text{int}} = u'/(\sqrt{1+u'^2}) z_\tau^{u,\text{int}} - 1/(\sqrt{1+u'^2}) z_\nu^{u,\text{int}}$. Since, by Theorem 2.3.1, $z_\tau^{u,\text{int}} = z_\tau^{u,\text{ext}}$ and (also, by Lemma 2.2.2) $z_\nu^{u,\text{int}} = g + z_\nu^{u,\text{ext}}$ on Γ_u, we have

$$z_y^{u,\text{int}}\big|_{\Gamma_u} = \left(\frac{u'}{\sqrt{1+u'^2}} z_\tau^{u,\text{ext}} - \frac{1}{\sqrt{1+u'^2}} \left[g + z_\nu^{u,\text{ext}} \right] \right) \Big|_{\Gamma_u} . \tag{2.3.38}$$

The corollary follows due to the Ω_u-*global* regularity of $z^{u,\text{ext}}$, and by the imbedding theorem. Indeed,

$$\begin{aligned} \|z_y^{u,\text{int}}\|_{C^{0,1-\frac{2}{q}}(\overline{\Gamma_u})} &= \left\| \frac{u'}{\sqrt{1+u'^2}} z_\tau^{u,\text{ext}} - \frac{1}{\sqrt{1+u'^2}} \left[g + z_\nu^{u,\text{ext}} \right] \right\|_{C^{0,1-\frac{2}{q}}(\overline{\Gamma_u})} \\ &\le c\|u\|_{H^3(-1,1)} \left[\|z^{u,\text{ext}}\|_{C^{1,1-\frac{2}{q}}(\overline{\Gamma_u})} + \|g\|_{C^{0,1-\frac{2}{q}}(\overline{\Gamma_u})} \right] \\ &\le c\|u\|_{H^3(-1,1)} \left[\|z^{u,\text{ext}}\|_{W^{2,q}(\Omega_u)} + \|g\|_{W^{1,q}(\Omega_u)} \right] \\ &\le c\left(\|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)} \right). \end{aligned} \tag{2.3.39}$$

We finish this section with the consideration of existence of a minimizer. Now, in order to claim existence of a minimizer, i.e., existence of a solution of the relaxed problem, one needs compactness. One way of introducing compactness would be to bound the set of admissible shapes to

$$U_b = \{u \in U; \|u\|_{H^3(-1,1)} \le b\}, \tag{2.3.40}$$

where b is some prescribed (large) positive constant.

PROPOSITION 2.3.1. *Let $g \in W^{1,q}(\Omega)$, for some $q > 2$. Then, there exists an $u^* \in U_b$ such that*

$$\Phi(u^*) = \min_{u \in U_b} \Phi(u). \tag{2.3.41}$$

Proof. Let $(u_n)_{n=1,2,\ldots} \subset U_b$ be a minimizing sequence. By Theorem 2.3.1 we know that

$$\|z^{u_n}\|_{H^1(\Omega)} + \|z^{u_n}\|_{C^{1-\frac{2}{q}}(\bar{\Omega})} \le c. \tag{2.3.42}$$

By taking subsequences, if necessary, we can assume without loss of generality that there exist $u^* \in U_b$ and $z^* \in H^1(\Omega)$ such that

$$u^n \to u^* \text{ in } H^2(-1,1), \tag{2.3.43}$$

$$z^{u_n} \rightharpoonup z^* \text{ weakly in } H^1(\Omega),\ z^{u_n} \to z^* \text{ in } C^0(\bar{\Omega}). \tag{2.3.44}$$

Recall that

$$-\int_\Omega \nabla z^{u_n} \cdot \nabla \varphi = \int_{\Gamma_{u_n}} g\varphi d\sigma \tag{2.3.45}$$

for all $\varphi \in H^1(\Omega)$ such that $\varphi|_{\{y=0\}} = \varphi|_{\{y=2\}} = 0$. If, in addition, $\varphi \in C^1(\bar{\Omega})$ then it is easy to see that

$$\lim_{n\to\infty} \int_{\Gamma_{u_n}} g\varphi d\sigma = \int_{\Gamma_{u^*}} g\varphi d\sigma. \tag{2.3.46}$$

Hence, for such φ we can pass $n \to \infty$ in (2.3.45) to conclude

$$-\int_{\Omega} \nabla z^* \cdot \nabla \varphi = \int_{\Gamma_{u^*}} g\varphi d\sigma \tag{2.3.47}$$

for any $\varphi \in C^1(\bar{\Omega})$ such that $\varphi|_{\{y=0\}} = \varphi|_{\{y=2\}} = 0$. But then, by the density, (2.3.47) holds for all $\varphi \in H^1(\Omega)$ such that $\varphi|_{\{y=0\}} = \varphi|_{\{y=2\}} = 0$. We conclude, by uniqueness, that $z^* = z^{u^*}$. Now since $\Phi(u_n) = \frac{1}{2}\int_{\Gamma_{u_n}} (z^{u_n})^2 d\sigma$, (2.3.43) and (2.3.44) imply that $\lim_{n\to\infty} \Phi(u_n) = \Phi(u^*)$, which completes the proof of the proposition.

2.4. Differentiability properties of the variational functional Φ. Since our problem is to minimize functional Φ, we want to derive information about the *multivalued* generalized gradient of Φ (see also Remark 2.4.1).

To make our results more precise, we shall introduce several definitions.

Let Φ be a real-valued function on the subset U of the Banach space X.

DEFINITION 2.4.1. *Φ is said to be* directionally differentiable *at $u \in U$ if the limit*

$$\lim_{\lambda \downarrow 0} \frac{\Phi(u+\lambda v) - \Phi(u)}{\lambda} \tag{2.4.1}$$

exists for any $v \in X$ such that $u + \lambda v \in U$, for small enough $\lambda > 0$. If that is the case, then the limit in (2.4.1) *is called* directional derivative *and it is denoted by $\Phi'(u; v)$.*

DEFINITION 2.4.2. *Φ is said to be* subdifferentiable *at u, if there exists an $f \in X^*$ such that*

$$\Phi'(u; v) \geq f(v) \tag{2.4.2}$$

for every $v \in X$ such that $u + \lambda v \in U$, for small enough $\lambda > 0$. Set of all such f's is called subdifferential, *and it is denoted by $\partial_* \Phi(u)$.*

DEFINITION 2.4.3. *Φ is said to be* superdifferentiable *at u, if there exists an $f \in X^*$ such that*

$$\Phi'(u; v) \leq f(v) \tag{2.4.3}$$

for every $v \in X$ such that $u + \lambda v \in U$, for small enough $\lambda > 0$. Set of all such f's is called superdifferential, *and it is denoted by $\partial^* \Phi(u)$.* If Φ is both sub- and superdifferentiable at $u \in \text{int}(U)$, and moreover $\partial_* \Phi(u) \cap \partial^* \Phi(u) \neq \emptyset$, then $\partial_* \Phi(u) \cap \partial^* \Phi(u)$ is a singleton and Φ is Gâteaux differentiable.

We go back now to our problem. Of course, $X = H_0^3(-1, 1)$, U is defined in (2.1.5).

Proceeding, define the adjoint variable p^u, as a solution of the (adjoint) equation

$$\begin{aligned} \Delta p^u &= \eta_u \text{ in } \Omega, \\ p^u &= 0 \text{ in } \{(x,0); -a < x < a\} \cup \{(x,2); -a < x < a\}, \\ p_x^u &= 0 \text{ in } \{(\pm a, y); 0 < y < 2\}, \end{aligned} \tag{2.4.4}$$

where $\eta_u \in H^{-1}(\Omega)$ is a (signed) measure given by

$$\eta_u(\varphi) = \int_{\Gamma_u} z^u \varphi d\sigma. \tag{2.4.5}$$

Obviously, (2.4.4) is the same type of equation as (2.2.2).

In this section, as before, $z^{u,\text{ext}} = z^u|_{\Omega_u}$ and $z^{u,\text{int}} = z^u|_{\Omega\setminus\Omega_u}$; also, later we shall use the notation $p^{u,\text{ext}} = p^u|_{\Omega_u}$ and $p^{u,\text{int}} = p^u|_{\Omega\setminus\Omega_u}$. That is essential in this calculation, since z^u and p^u are *not* differentiable *across* the Γ_u.

LEMMA 2.4.1. *Let $g \in W^{2,q}(\Omega)$, for some $q \geq 2$. Then*

$$p^{u,\text{ext}} \in W^{2,q}(\Omega_u),\ p^{u,\text{int}} \in W^{2,q}_{\{y>0\}-loc}(\Omega\setminus\bar{\Omega}_u) \tag{2.4.6}$$

and the a priori estimates

$$\|p^{u,\text{ext}}\|_{W^{2,q}(\Omega_u)} \leq c\left(\|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right), \tag{2.4.7}$$

and

$$\begin{aligned}&\|p^{u,\text{int}}\|_{W^{2,q}((\Omega\setminus\Omega_u)\cap\{y\geq\epsilon\})}\\ &\quad\leq c\left(\epsilon, \|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right)\end{aligned} \tag{2.4.8}$$

hold.

Proof. Comparing (2.2.2) and (2.4.4) we see that the only difference is in right-hand sides. Namely, in (2.4.5), $z^u \notin W^{2,q}(\Omega)$. Nevertheless, for example, $z^{u,\text{ext}} \in W^{2,q}(\Omega_u)$, and since η_u depends on z^u only through the trace on Γ_u, and since z^u and $z^{u,\text{ext}}$ have same traces on Γ_u we easily conclude the proof of the lemma.

We shall use the usual notation: $v^+ \stackrel{\text{def}}{=} v\mathbf{I}_{\{v>0\}}$, and $v^- \stackrel{\text{def}}{=} -v\mathbf{I}_{\{v<0\}}$. So, $v = v^+ - v^-$.

Now we are ready to state the following.

THEOREM 2.4.1. *Let $g \in W^{2,q}(\Omega)$, for some $q > 2$. Then Φ is directionally differentiable at any $u \in U$ such that $u(x) > 0$ for $-1 < x < 1$, and*

$$\begin{aligned}&\Phi'(u;v)\\ &= \int_{-1}^{1}\left(z^u(z_y^{u,\text{ext}}v^+ - z_y^{u,\text{int}}v^-)\sqrt{1+u'^2} + \frac{1}{2}(z^u)^2\frac{u'v'}{\sqrt{1+u'^2}}\right)dx\\ &\quad + \int_{\Gamma_u}\left((gp^{u,\text{ext}})_y v^+ - (gp^{u,\text{int}})_y v^-\right)d\sigma + \int_{-1}^{1} gp^u\frac{u'v'}{\sqrt{1+u'^2}}dx.\end{aligned} \tag{2.4.9}$$

Moreover, if

$$z^u z_y^{u,\text{int}} + (gp^{u,\text{int}})_y \leq z^u z_y^{u,\text{ext}} + (gp^{u,\text{ext}})_y \ \ a.e.\ in\ (-1,1), \tag{2.4.10}$$

then Φ is subdifferentiable at u and

$$\begin{aligned}&\partial_*\Phi(u)\\ &= \left[\left(z^u z_y^{u,\text{int}} + (gp^{u,\text{int}})_y\right)\sqrt{1+u'^2}, \left(z^u z_y^{u,\text{ext}} + (gp^{u,\text{ext}})_y\right)\sqrt{1+u'^2}\right]\\ &\quad - \left(\frac{u'}{\sqrt{1+u'^2}}\left(\frac{1}{2}(z^u)^2 + gp^u\right)\right)'\\ &\stackrel{\text{def}}{=} [l\partial_*\Phi(u), r\partial_*\Phi(u)] \subset L^\infty(-1,1).\end{aligned} \tag{2.4.11}$$

On the other hand, if

$$z^u z_y^{u,\text{int}} + (gp^{u,\text{int}})_y \geq z^u z_y^{u,\text{ext}} + (gp^{u,\text{ext}})_y \ \ a.e.\ in\ (-1,1), \tag{2.4.12}$$

then Φ is superdifferentiable at u and

$$\begin{aligned}\partial^*\Phi(u) &= \left[\left(z^u z_y^{u,\mathrm{ext}} + (gp^{u,\mathrm{ext}})_y\right)\sqrt{1+u'^2}, \left(z^u z_y^{u,\mathrm{int}} + (gp^{u,\mathrm{int}})_y\right)\sqrt{1+u'^2}\right] \\ &\quad - \left(\frac{u'}{\sqrt{1+u'^2}}\left(\frac{1}{2}(z^u)^2 + gp^u\right)\right)' \\ &\stackrel{\mathrm{def}}{=} [l\partial^*\Phi(u), r\partial^*\Phi(u)] \subset L^\infty(-1,1). \end{aligned} \tag{2.4.13}$$

Proof. We attempt to differentiate Φ. To this end, for given $u \in U$ and a suitable direction $v \in H_0^3(-1,1)$ (suitable in a sense that $u + \lambda v \in U$ for small enough $\lambda > 0$) we try to compute the (one-sided) directional derivative $\Phi'(u;v)$. Using the regularity result (Theorem 2.3.1, and Corollary 2.3.2), we compute

$$\begin{aligned}\Phi'(u;v) &= \lim_{\lambda \downarrow 0} \frac{\Phi(u+\lambda v) - \Phi(u)}{\lambda} \\ &= \lim_{\lambda\downarrow 0} \frac{1}{2\lambda}\left(\int_{\Gamma_{u+\lambda v}} (z^{u+\lambda v})^2 d\sigma - \int_{\Gamma_u} (z^u)^2 d\sigma\right) \\ &= \int_{-1}^{1}\left(z^u(z_y^{u,\mathrm{ext}} v^+ - z_y^{u,\mathrm{int}} v^-)\sqrt{1+u'^2} + \frac{1}{2}(z^u)^2 \frac{u'v'}{\sqrt{1+u'^2}}\right) dx \\ &\quad + \lim_{\lambda\downarrow 0} \frac{1}{2\lambda}\int_{\Gamma_u}\left((z^{u+\lambda v})^2 - (z^u)^2\right) d\sigma, \end{aligned} \tag{2.4.14}$$

Before proceeding with the proof, we shall need the following lemma (more precisely, its corollary).

LEMMA 2.4.2. *Under previous assumptions on u, and v, and for any $\alpha < 1$ the following estimate holds*:

$$\|z^{u+\lambda v} - z^u\|_{C^0(\bar{\Omega})} \le c\lambda^\alpha. \tag{2.4.15}$$

Proof. We need to compare $z^{u+\lambda v}$ and z^u. This is difficult to do in the original domain Ω since the (singular) right-hand sides of the equations that they satisfy act on disjoint sets, so that there is no obvious cancellation. So, the idea of the proof is to map the original domain into different domains in such a way that the cancellation takes place.

As before, let Ξ_u be the map with the image Ω given by the formula $\Xi_u(x,y) = (x, y + u(x))$. Then $\Xi_u^{-1}(x,y) = (x, y - u(x))$, and (set $A = (x,y)$) $\mathrm{dist}(\Xi_{u+\lambda v}^{-1}(A) - \Xi_u^{-1}(A)) \le c\lambda$. Now consider $\tilde{z}^{u+\lambda v}$ and $\tilde{z}^u$ defined as

$$\tilde{z}^{u+\lambda v} = z^{u+\lambda v} \circ \Xi_{u+\lambda v},\ \tilde{z}^u = z^u \circ \Xi_u, \tag{2.4.16}$$

and operators L_u and $L_{u+\lambda v}$ defined by

$$L_u w = \Delta w + w_{yy}(u_x)^2 - 2w_{xy}u_x - w_y u_{xx}, \tag{2.4.17}$$

$$\begin{aligned} L_{u+\lambda v} w &= \Delta w + w_{yy}(u_x + \lambda v_x)^2 - 2w_{xy}(u_x + \lambda v_x) - w_y(u_{xx} + \lambda v_x) \\ &= L_u w + \lambda\left[w_{yy}(2u_x v_x + \lambda v_x^2) - 2w_{xy}v_x - w_y v_{xx}\right]. \end{aligned} \tag{2.4.18}$$

Then $\tilde{z}^{u+\lambda v} - \tilde{z}^u$ satisfies the equation

$$L_u\left(\tilde{z}^{u+\lambda v} - \tilde{z}^u\right) = \gamma - \lambda\left[\tilde{z}_{yy}^{u+\lambda v}(2u_x v_x + \lambda v_x^2) - 2\tilde{z}_{xy}^{u+\lambda v} v_x - \tilde{z}_y^{u+\lambda v} v_{xx}\right] \tag{2.4.19}$$

in $\Xi_{u+\lambda v}^{-1}(\Omega) \cap \Xi_u^{-1}(\Omega)$, where

$$\gamma(\varphi) \stackrel{\text{def}}{=} \int_{\{y=0\}} (G_1 - G_2)\,\varphi dx \tag{2.4.20}$$

and where

$$\begin{aligned} G_1(x,y) &\stackrel{\text{def}}{=} g(x, y + u(x) + \lambda v(x))\sqrt{1 + (u'(x) + \lambda v'(x))^2}, \\ G_2(x,y) &\stackrel{\text{def}}{=} g(x, y + u(x))\sqrt{1 + (u'(x))^2}. \end{aligned} \tag{2.4.21}$$

Observe that

$$\|G_1 - G_2\|_{W^{1,q}\left(\Xi_{u+\lambda v}^{-1}(\Omega) \cap \Xi_u^{-1}(\Omega)\right)} \leq c\lambda. \tag{2.4.22}$$

Now since

$$\operatorname{dist}\left(\partial\left(\Xi_{u+\lambda v}^{-1}(\Omega)\right), \partial\left(\Xi_u^{-1}(\Omega)\right)\right) \leq c\lambda \tag{2.4.23}$$

and because of the Hölder continuity of $z^{u+\lambda v}$ and z^u, we conclude that

$$\|\tilde{z}^{u+\lambda v} - \tilde{z}^u\|_{C^0\left(\partial\left(\Xi_{u+\lambda v}^{-1}(\Omega) \cap \Xi_u^{-1}(\Omega)\right)\right)} \leq c\lambda^\alpha. \tag{2.4.24}$$

Then (2.4.19), (2.4.22), and (2.4.24) imply that

$$\|\tilde{z}^{u+\lambda v} - \tilde{z}^u\|_{C^0\left(\overline{\Xi_{u+\lambda v}^{-1}(\Omega) \cap \Xi_u^{-1}(\Omega)}\right)} \leq c\lambda^\alpha. \tag{2.4.25}$$

Then we have (set $A = (x, y)$)

$$\begin{aligned} &\left|z^{u+\lambda v}(A) - z^u(A)\right| \\ &\quad = \left|\tilde{z}^{u+\lambda v}\left(\Xi_{u+\lambda v}^{-1}(A)\right) - \tilde{z}^u\left(\Xi_u^{-1}(A)\right)\right| \\ &\quad \leq \left|\tilde{z}^{u+\lambda v}\left(\Xi_{u+\lambda v}^{-1}(A)\right) - \tilde{z}^u\left(\Xi_{u+\lambda v}^{-1}(A)\right)\right| \\ &\qquad + \left|\tilde{z}^u\left(\Xi_{u+\lambda v}^{-1}(A)\right) - \tilde{z}^u\left(\Xi_u^{-1}(A)\right)\right| \\ &\quad \leq c\lambda^\alpha + c\lambda^\alpha = c\lambda^\alpha. \end{aligned} \tag{2.4.26}$$

In (2.4.26), we also used the Hölder continuity of $\tilde{z}^u$. This completes the proof of the lemma.

COROLLARY 2.4.1.

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \int_{\Gamma_u} \left(z^{u+\lambda v} - z^u\right)^2 d\sigma = 0. \tag{2.4.27}$$

Proof. Take $\alpha > \frac{1}{2}$ in the lemma. Then

$$\frac{\left\|z^{u+\lambda v} - z^u\right\|_{C^0(\bar{\Omega})}^2}{\lambda} \leq c\lambda^\beta, \ \beta = 2\alpha - 1 > 0. \tag{2.4.28}$$

Now we can proceed with the proof of the theorem. We compute the last term in (2.4.14).

$$\begin{aligned}
&\lim_{\lambda\downarrow 0}\frac{1}{2\lambda}\int_{\Gamma_u}\left((z^{u+\lambda v})^2-(z^u)^2\right)d\sigma\\
&=\lim_{\lambda\downarrow 0}\frac{1}{\lambda}\int_{\Gamma_u}(z^{u+\lambda v}-z^u)z^u d\sigma+\lim_{\lambda\downarrow 0}\frac{1}{2\lambda}\int_{\Gamma_u}\left(z^{u+\lambda v}-z^u\right)^2 d\sigma\\
&=\lim_{\lambda\downarrow 0}\frac{1}{\lambda}\int_{\Gamma_u}(z^{u+\lambda v}-z^u)z^u d\sigma=-\lim_{\lambda\downarrow 0}\frac{1}{\lambda}\int_{\Omega}\nabla p^u\cdot\nabla(z^{u+\lambda v}-z^u)\\
&=\lim_{\lambda\downarrow 0}\frac{1}{\lambda}\left(\int_{\Gamma_{u+\lambda v}}gp^u d\sigma-\int_{\Gamma_u}gp^u d\sigma\right)\\
&=\int_{\Gamma_u}\left((gp^{u,\mathrm{ext}})_y v^+-(gp^{u,\mathrm{int}})_y v^-\right)d\sigma+\int_{-1}^{1}gp^u\frac{u'v'}{\sqrt{1+u'^2}}dx. \qquad (2.4.29)
\end{aligned}$$

Now from (2.4.14) and (2.4.29) we conclude that Φ is directionally differentiable, and that (2.4.9) holds. Furthermore, if (2.4.10) holds, then

$$\begin{aligned}
&\Phi'(u;v)\\
&=\int_{-1}^{1}\left(z^u(z_y^{u,\mathrm{ext}}v^+-z_y^{u,\mathrm{int}}v^-)\sqrt{1+u'^2}+\frac{1}{2}(z^u)^2\frac{u'v'}{\sqrt{1+u'^2}}\right)dx\\
&\quad+\int_{\Gamma_u}\left((gp^{u,\mathrm{ext}})_y v^+-(gp^{u,\mathrm{int}})_y v^-\right)d\sigma+\int_{-1}^{1}gp^u\frac{u'v'}{\sqrt{1+u'^2}}dx\\
&\geq\int_{-1}^{1}\left(\tau-\left(\frac{u'}{\sqrt{1+u'^2}}\left(\frac{1}{2}(z^u)^2+gp^u\right)\right)'\right)v dx \qquad (2.4.30)
\end{aligned}$$

for all

$$\begin{aligned}
\tau\in\Big[&\left(z^u z_y^{u,\mathrm{int}}+(gp^{u,\mathrm{int}})_y\right)\sqrt{1+u'^2},\\
&\left(z^u z_y^{u,\mathrm{ext}}+(gp^{u,\mathrm{ext}})_y\right)\sqrt{1+u'^2}\Big]. \qquad (2.4.31)
\end{aligned}$$

This proves that Φ is subdifferentiable at u and that (2.4.11) holds. Similarly, one can consider superdifferentiability of Φ. So the theorem follows.

REMARK 2.4.1. *The previous suggests the numerical algorithm* (the steepest descent method) *for minimization of* Φ, *i.e., for the numerical solution of the relaxed free boundary problem:*

Choose $u_0\in U$. *If* $u_n\in U$ *is already known, then* u_{n+1} *is determined as follows:*

- *compute* z^{u_n} *as a solution of* (2.2.2);
- *compute* p^{u_n} *as a solution of* (2.4.4);
- *if* (2.4.10) *holds, compute an* u_{n+1} *such that*

$$u_{n+1}\in\left(u_n-\rho_n A^{-1}\left(\partial_*\Phi(u_n)\right)\right)\cap U,\ \rho_n>0, \qquad (2.4.32)$$

and if (2.4.12) *holds, compute an* u_{n+1} *such that*

$$u_{n+1}\in\left(u_n-\rho_n A^{-1}\left(\partial^*\Phi(u_n)\right)\right)\cap U,\ \rho_n>0. \qquad (2.4.33)$$

Here, A is the isomorphism between $H_0^3(-1,1)$ and its dual. More precisely, $\bar{u} = A^{-1}(l)$ means that $\bar{u}$ solves the following boundary value problem:

$$-\frac{d^6\bar{u}}{dx^6} = l \text{ in } (-1,1),$$

$$\bar{u}(-1) = \bar{u}'(-1) = \bar{u}''(-1) = \bar{u}(1) = \bar{u}'(1) = \bar{u}''(1) = 0. \tag{2.4.34}$$

If neither (2.4.10) *nor* (2.4.12) *holds, i.e., if Φ is neither convex nor concave at the point u_n, then it is more delicate to determine the steep(est) descent direction.*

3. Stokes flow.

3.1. Statement of the problem. The purpose of this section is to extend the previous results to the case of Stokes flow.

We consider a motionless body $\mathcal{B}$ in a viscous incompressible fluid moving in a bounded region Λ containing $\mathcal{B}$. The boundary of the region Λ will be denoted as $\partial\Lambda$. Fluid is moving at the velocity $\mathbf{h}$ at $\partial\Lambda$, and $\mathbf{h}$ is such that $\int_{\partial\Lambda} \mathbf{h}\cdot\mathbf{n}d\sigma = 0$, where $\mathbf{n}$ is the unit (exterior) normal to $\partial\Lambda$.

The boundary of the body $\partial\mathcal{B}$ consists of two disjoint and connected parts Σ and Γ, $\partial\mathcal{B} = \Sigma\cup\Gamma$. We shall suppose that Γ can be described as

$$\Gamma = \Gamma_u = \{(x_1, u(x_1)); -1 < x_1 < 1\} \tag{3.1.1}$$

for some function $u \in U$, where

$$U = \left\{u \in H_0^3(-1,1); 0 \le u(x_1) \le 1, -1 < x_1 < 1\right\}. \tag{3.1.2}$$

So, if we want to emphasize the dependence on $u \in U$ we shall write also $\mathcal{B} = \mathcal{B}_u$.

Denote by Ω_u the actual flow region $\Omega_u \overset{\text{def}}{=} \Lambda \setminus \bar{\mathcal{B}}_u$. Also, we assume that Σ is such that $\partial\mathcal{B}$ is sufficiently regular. Finally, denote,

$$\Omega = \{(x_1,x_2);\ -1 < x_1 < 1,\ 0 < x_2 < u(x_1)\}\cup\Omega_u\cup\Gamma_u, \tag{3.1.3}$$

so that $\Omega\setminus\bar{\Omega}_u = \{(x_1,x_2);\ -1 < x_1 < 1,\ 0 < x_2 < u(x_1)\}$.

Now, the velocity vector field of the fluid $\mathbf{w} = \mathbf{w}^u$, and the pressure p, are the solution of the Stokes system

$$\begin{gathered} -\nu\Delta\mathbf{w} + \nabla p = \mathbf{0} \text{ in } \Omega_u,\ \nabla\cdot\mathbf{w} = 0 \text{ in } \Omega_u, \\ \mathbf{w} = 0 \text{ in } \Gamma_u\cup\Sigma,\ \mathbf{w} = \mathbf{h} \text{ in } \partial\Lambda \end{gathered} \tag{3.1.4}$$

We observe that the pressure p in (3.1.4) is determined only uniquely up to the additive constant.

The problem we propose is the following:

For given $\mathbf{g} = \mathbf{g}(x_1,x_2)$ such that (we will not always have to assume this much)

$$\mathbf{g} \in C^{1,1}(\Omega)^2, \tag{3.1.5}$$

$$\mathbf{g} = \mathbf{0} \text{ in } \Omega\cap\{|x_1| > 1\}, \tag{3.1.6}$$

find (if possible) $u \in U$ such that if $\mathbf{w}^u$ is the corresponding solution of (3.1.4), then also

$$-pn_j + \nu\left(\frac{\partial w_j}{\partial x_i} + \frac{\partial w_i}{\partial x_j}\right)n_i = g_j \text{ in } \Gamma_u,\ j = 1,2, \tag{3.1.7}$$

where $\mathbf{n} = (n_i)_{i=1,2}$ is the unit normal exterior to Ω_u. We observe now that if (3.1.7) is to be satisfied in addition to (3.1.4) (and if (3.1.7) and (3.1.4) have a solution) then pressure p is determined *uniquely*. Also, we note that condition (3.1.7) means that fluid motion exhibits force distribution $\mathbf{g}$ on the boundary Γ_u. So, the problem we propose is to find a shape of the immersed body so that the prescribed force field is generated at the boundary.

We can simplify this problem right away. Let, as usual, $V = V(\Omega) \stackrel{\text{def}}{=} \{\mathbf{u} \in H_0^1(\Omega)^2;\ \nabla \cdot \mathbf{u} = 0\}$. We have the following lemma.

LEMMA 3.1.1. *If* $\mathbf{w} \in V \cap H^2(\Omega)^2$, *and* $\partial\Omega \in C^{0,1}$, *then*

$$\frac{\partial w_i}{\partial x_j} n_i = 0 \text{ on } \partial\Omega. \tag{3.1.8}$$

Proof. Let $\varphi \in C^\infty(R^2)^2$ be an arbitrary function. We have

$$0 = \int_\Omega \frac{\partial w_i}{\partial x_i} \frac{\partial \varphi_j}{\partial x_j} = \int_\Omega \frac{\partial w_i}{\partial x_j} \frac{\partial \varphi_j}{\partial x_i} = \int_{\partial\Omega} \frac{\partial w_i}{\partial x_j} n_i \varphi_j. \tag{3.1.9}$$

So the lemma follows.

Lemma 3.1.1 implies that requesting (3.1.7) (in addition to (3.1.4)) is equivalent to requesting

$$-pn_j + \nu \frac{\partial w_j}{\partial x_i} n_i = g_j \text{ in } \Gamma_u,\ j = 1, 2. \tag{3.1.10}$$

Note that (3.1.10) is closely related to the equvalence of problems (3.2.2) and (3.2.3) (and hence, (3.2.4)).

3.2. Relaxation of the problem. Suppose that there exist a u, and a pair $(\mathbf{w}^u, p^u)$, a solution of (3.1.4), such that also (3.1.7) holds. So, we suppose existence for the free boundary problem (3.1.4) and (3.1.7). To refer to such an assumption we shall say that u is supposed to be an *exact shape*. Extend $\mathbf{w}^u$ from Ω_u to Ω as $\mathbf{z}^u$:

$$\mathbf{z}^u = \begin{cases} 0 & \text{on } \Omega \setminus \bar{\Omega}_u, \\ \mathbf{w}^u & \text{on } \Omega_u. \end{cases} \tag{3.2.1}$$

Lemma 3.2.1 follows.

LEMMA 3.2.1. *If* u *is an exact shape, then* $\mathbf{z}^u \in H^1(\Omega)^2$, *and it solves the Stokes system* (*with singular right-hand side*)

$$\frac{\nu}{2} \int_\Omega D(\mathbf{z}^u) : D(\varphi) = \xi_u(\varphi),\ \ \forall \varphi \in V, \tag{3.2.2}$$

i.e.,

$$\nu \int_\Omega \nabla \mathbf{z}^u : \nabla \varphi = \xi_u(\varphi),\ \ \forall \varphi \in V, \tag{3.2.3}$$

i.e.,

$$\begin{aligned} -\nu \Delta \mathbf{z}^u + \nabla p &= \xi_u,\ \ \textit{in } \mathcal{D}'(\Omega)^2, \\ \nabla \cdot \mathbf{z}^u &= 0,\ \ \textit{a.e. in } \Omega, \end{aligned} \tag{3.2.4}$$

where $\xi_u \in H^{-1}(\Omega)^2$ *is a signed vector measure given by*

$$\xi_u(\varphi) = \int_{\Gamma_u} \mathbf{g} \cdot \varphi d\sigma. \tag{3.2.5}$$

Here, $D(\mathbf{z}^u) : D(\varphi^u) = (\frac{\partial z_j}{\partial x_i} + \frac{\partial z_i}{\partial x_j})(\frac{\partial \varphi_j}{\partial x_i} + \frac{\partial \varphi_i}{\partial x_j})$, and throughout the paper the summation convention is assumed.

Proof of the Lemma. The proof is similar to the proof of Lemma 2.2.1; see also the proof of Lemma 3.2.2.

LEMMA 3.2.2. *Let* $u \in U$ *be given. Then, if* $\mathbf{z}^u$ *is a solution of* (3.2.2) *and if it happens that* $\mathbf{z}^u|_{\Gamma_u} = 0$, *then there exists pressure* p^u *such that* $(\mathbf{z}^u|_{\Omega_u}, p^u|_{\Omega_u})$ *is a solution of* (3.2.2) *and* (3.1.7), *i.e.,* u *is an exact shape.*

Proof. Observe that any test function $\varphi \in V$ must satisfy $0 = \int_{\Omega\setminus\bar{\Omega_u}} \nabla \cdot \varphi = \int_{\partial(\Omega\setminus\bar{\Omega_u})} \varphi \cdot \mathbf{n} d\sigma = \int_{\Gamma_u} \varphi \cdot \mathbf{n} d\sigma$. From (3.2.2) we see (here we assume the regularity of $\mathbf{z}^u|_{\Omega\setminus\bar{\Omega_u}}$ and $\mathbf{z}^u|_{\Omega_u}$, to be proved in the next section),

$$\begin{aligned}\int_{\Gamma_u} \mathbf{g} \cdot \varphi d\sigma &= \frac{\nu}{2}\left(\int_{\Omega\setminus\bar{\Omega_u}} D(\mathbf{z}^u) : D(\varphi) + \int_{\Omega_u} D(\mathbf{z}^u) : D(\varphi)\right) \\ &= \int_{\Omega\setminus\bar{\Omega_u}} (-\nu\Delta\mathbf{z}^u)\cdot\varphi + \int_{\Omega_u} (-\nu\Delta\mathbf{z}^u)\cdot\varphi \\ &\quad + \int_{\partial(\Omega\setminus\bar{\Omega_u})} \nu D(\mathbf{z}^u)\mathbf{n}\cdot\varphi d\sigma + \int_{\partial\Omega_u} \nu D(\mathbf{z}^u)\mathbf{n}\cdot\varphi d\sigma. \end{aligned} \tag{3.2.6}$$

To fix ideas, let $\mathbf{n}$ be the unit normal exterior to Ω_u. Then (3.2.6) is equal to

$$\begin{aligned}&\int_{\Omega\setminus\bar{\Omega_u}} (-\nu\Delta\mathbf{z}^u)\cdot\varphi + \int_{\Omega_u} (-\nu\Delta\mathbf{z}^u)\cdot\varphi \\ &+ \int_{\Gamma_u} \nu\left(D(\mathbf{z}^u)^{\text{ext}} - D(\mathbf{z}^u)^{\text{int}}\right)\mathbf{n}\cdot\varphi d\sigma \end{aligned} \tag{3.2.7}$$

where $f^{\text{int}} \overset{\text{def}}{=} f|_{\Omega\setminus\bar{\Omega_u}}$ and $f^{\text{ext}} \overset{\text{def}}{=} f|_{\Omega_u}$.

We conclude that there exists $p^u \in L^2(\Omega)$ such that

$$-\nu\Delta\mathbf{z}^u + \nabla p^u = \mathbf{0} \text{ in } \Omega\setminus\bar{\Omega}_u, \tag{3.2.8}$$

$$-\nu\Delta\mathbf{z}^u + \nabla p^u = \mathbf{0} \text{ in } \Omega_u. \tag{3.2.9}$$

For such p we have

$$\begin{aligned}\int_{\Gamma_u} \mathbf{g}\cdot\varphi d\sigma &= \int_{\Omega\setminus\bar{\Omega_u}} (-\nabla p^u)\cdot\varphi + \int_{\Omega_u} (-\nabla p^u)\cdot\varphi \\ &\quad + \int_{\Gamma_u} \nu\left(D(\mathbf{z}^u)^{\text{ext}} - D(\mathbf{z}^u)^{\text{int}}\right)\mathbf{n}\cdot\varphi d\sigma \\ &= \int_\Omega p^u \nabla\cdot\varphi - \int_{\partial(\Omega\setminus\bar{\Omega_u})} p^u\varphi\cdot\mathbf{n}d\sigma - \int_{\partial\Omega_u} p^u\varphi\cdot\mathbf{n}d\sigma \\ &\quad + \int_{\Gamma_u} \nu\left(D(\mathbf{z}^u)^{\text{ext}} - D(\mathbf{z}^u)^{\text{int}}\right)\mathbf{n}\cdot\varphi d\sigma. \end{aligned} \tag{3.2.10}$$

Let $[f]_{\text{int}}^{\text{ext}} \overset{\text{def}}{=} f^{\text{ext}}|_{\Gamma_u} - f^{\text{int}}|_{\Gamma_u}$ be the jump accross Γ_u. Then (3.2.10) implies that

$$\int_{\Gamma_u} g_j \varphi_j d\sigma = \int_{\Gamma_u} \left[-p^u n_j + \nu \left(\frac{\partial z_j^u}{\partial x_i} + \frac{\partial z_i^u}{\partial x_j} \right) n_i \right]_{\text{int}}^{\text{ext}} \varphi_j d\sigma \tag{3.2.11}$$

for all φ such that $\int_{\Gamma_u} \varphi_j n_j = 0$. This implies

$$\left[-p^u n_j + \nu \left(\frac{\partial z_j^u}{\partial x_i} + \frac{\partial z_i^u}{\partial x_j} \right) n_i \right]_{\text{int}}^{\text{ext}} - g_j = (\text{const}) n_j, \ j = 1, 2. \tag{3.2.12}$$

We note that (3.2.12) holds for any p^u such that (3.2.8) and (3.2.9) hold.

Now suppose that $\mathbf{z}^u|_{\Gamma_u} = 0$. Then $\mathbf{z}^u|_{\Omega \setminus \bar{\Omega}_u} = 0$, and without loss of generality we can choose $p^u|_{\Omega \setminus \bar{\Omega}_u} = 0$. Then (3.2.12) implies

$$\left(-p^u n_j + \nu \left(\frac{\partial z_j^u}{\partial x_i} + \frac{\partial z_i^u}{\partial x_j} \right) n_i \right)^{\text{ext}} \Bigg|_{\Gamma_u} - g_j = (\text{const}) n_j, \ j = 1, 2, \tag{3.2.13}$$

for any p^u satisfying (3.2.9). So, there exists p^u such that (3.2.9) holds and such that

$$\left(-p^u n_j + \nu \left(\frac{\partial z_j^u}{\partial x_i} + \frac{\partial z_i^u}{\partial x_j} \right) n_i \right)^{\text{ext}} \Bigg|_{\Gamma_u} - g_j = 0, \ j = 1, 2, \tag{3.2.14}$$

which is nothing but (3.1.7).

Lemma 3.2.2 motivates the following definition.

DEFINITION 3.2.1. *$u^* \in U$ is said to solve the* relaxed free boundary problem *if the corresponding $\mathbf{z}^u$ defined by* (3.2.2), *is such that if*

$$\Phi(u) = \frac{1}{2} \int_{\Gamma_u} |\mathbf{z}^u|^2 d\sigma \tag{3.2.15}$$

then

$$\Phi(u^*) = \min_{u \in U} \Phi(u). \tag{3.2.16}$$

3.3. The state equation. Let s be the segment

$$s = \{-1 \leq x \leq 1, y = 0\}, \tag{3.3.1}$$

and let

$$S_\epsilon = \{(x, y); \text{dist}((x, y), s) \leq \epsilon\}. \tag{3.3.2}$$

Define

$$W^{2,q}_{s-\text{loc}}(\Omega) \overset{\text{def}}{=} \cap_{\epsilon > 0} W^{2,q}(\Omega \setminus S_\epsilon). \tag{3.3.3}$$

Now we have the next theorem.

THEOREM 3.3.1. *Let q be such that $1 < q < \infty$, and $\partial\Omega \in C^2$. Let $u \in U$. Then the state equation* (3.2.2) *has a unique weak solution $\mathbf{z}^u$, and*

$$\mathbf{z}^u \in V \cap C^\infty(\Omega \setminus \Gamma_u)^2. \tag{3.3.4}$$

More importantly, the following regularity results hold:
(a) L^p-estimate: *If* $\mathbf{g} \in W^{1,q}(\Omega)^2$, *then*

$$\begin{aligned} &\|\mathbf{z}^u\|_{W^{1,q}(\Omega)^2} + \|p^u\|_{L^q(\Omega)/R} \le c\left(1 + \|u\|_{C^{0,1}(-1,1)}\right) \\ &\cdot \left(\|\mathbf{g}\|_{W^{1,q}(\Omega)^2} + \|\mathbf{h}\|_{W^{1-\frac{1}{q},q}(\partial\Omega)^2}\right). \end{aligned} \tag{3.3.5}$$

(b) maximum modulus estimate: *If, moreover,* $q > 2$, *then*

$$\begin{aligned} &\|\mathbf{z}^u\|_{L^\infty(\Omega)^2} \le c\left(1 + \|u\|_{C^{0,1}(-1,1)}\right) \\ &\left(\|\mathbf{g}\|_{W^{1,q}(\Omega)^2} + \|\mathbf{h}\|_{L^\infty(\partial\Omega)^2}\right). \end{aligned} \tag{3.3.6}$$

(c) *If* $\mathbf{g} \in W^{2,q}(\Omega)^2$, *and* (3.1.6) *holds, then*

$$\mathbf{z}^{u,\mathrm{ext}} \in W^{2,q}(\Omega_u)^2,\ \mathbf{z}^{u,\mathrm{int}} \in W^{2,q}_{s-\mathrm{loc}}(\Omega \setminus \bar{\Omega}_u)^2 \tag{3.3.7}$$

and the a priori estimates

$$\begin{aligned} &\|\mathbf{z}^{u,\mathrm{ext}}\|_{W^{2,q}(\Omega_u)^2} + \|p^{u,\mathrm{ext}}\|_{W^{1,q}(\Omega)/R} \\ &\le c\left(\|u\|_{H^3(-1,1)}, \|\mathbf{g}\|_{W^{2,q}(\Omega)^2}, \|\mathbf{h}\|_{W^{2-\frac{1}{q},q}(\partial\Omega)^2}\right), \end{aligned} \tag{3.3.8}$$

and

$$\begin{aligned} &\|\mathbf{z}^{u,\mathrm{int}}\|_{W^{2,q}((\Omega\setminus\Omega_u)\cap\{y>\epsilon\})^2} + \|p^{u,\mathrm{int}}\|_{W^{1,q}((\Omega\setminus\Omega_u)\cap\{y>\epsilon\})/R} \\ &\le c\left(\epsilon, \|u\|_{H^3(-1,1)}, \|\mathbf{g}\|_{W^{2,q}(\Omega)^2}, \|\mathbf{h}\|_{W^{2-\frac{1}{q},q}(\partial\Omega)^2}\right) \end{aligned} \tag{3.3.9}$$

hold.

Proof. The interior regularity $\mathbf{z}^u \in C^\infty(\Omega \setminus \Gamma_u)^2$ follows easily (see, e.g., [6]). The proof of (3.3.5) is similar to the proof of (2.3.4). One has to use (see [5]) the following L^q-estimate for the Stokes problem:

$$\begin{aligned} &\|\mathbf{z}^u\|_{W^{1,q}(\Omega)^2} + \|p^u\|_{L^q(\Omega)/R} \\ &\le c\left\{\|F\|_{W^{-1,q}(\Omega)^2} + \|\mathbf{h}\|_{W^{1-\frac{1}{q},q}(\partial\Omega)^2}\right\}, \end{aligned} \tag{3.3.10}$$

where F is the right-hand side; in our case $F(\varphi) = \int_{\Gamma_u} \mathbf{g} \cdot \varphi d\sigma$.

We prove now (3.3.6). We use the following important result (see [13]) for biharmonic functions:

If $\Delta^2\varphi = 0$ in Ω, then

$$\|\nabla\varphi\|_{L^\infty(\Omega)^2} \le c\|\nabla\varphi\|_{L^\infty(\partial\Omega)^2}. \tag{3.3.11}$$

Let $\mathbf{w}$ solve the homogeneous Stokes problem

$$-\nu\Delta\mathbf{w} + \nabla p = 0 \text{ in } \Omega, \tag{3.3.12}$$

$$\nabla \cdot \mathbf{w} = 0 \text{ in } \Omega, \tag{3.3.13}$$

$$\mathbf{w} = \mathbf{h} \text{ on } \partial\Omega, \tag{3.3.14}$$

where $\mathbf{h}$ is such that $\int_{\partial\Omega} \mathbf{h} \cdot \mathbf{n} d\sigma = 0$. It is well known (see, e.g., [10]) that (3.3.13) and (3.3.14) imply the existence of φ such that $\mathbf{w} = \mathbf{curl}\varphi$. Here $\mathbf{curl}\varphi = \langle \frac{\partial\varphi}{\partial y}, -\frac{\partial\varphi}{\partial x} \rangle$.

Also let $\mathrm{curl}\mathbf{v} = \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y}$. Now, since curl $\mathbf{curl}\varphi = -\Delta\varphi$, and since $\mathrm{curl}\nabla p = 0$, we conclude from (3.3.12) that $\Delta^2\varphi = 0$, and hence (3.3.11) follows. But then we have

$$\begin{aligned} &\| \mathbf{w}\|_{L^\infty(\Omega)^2} = \|\mathbf{curl}\varphi\|_{L^\infty(\Omega)^2} = \|\nabla\varphi\|_{L^\infty(\Omega)^2} \\ &\leq c\|\nabla\varphi\|_{L^\infty(\partial\Omega)^2} = c\|\mathbf{curl}\varphi\|_{L^\infty(\partial\Omega)^2} = c\|\mathbf{h}\|_{L^\infty(\partial\Omega)^2}. \end{aligned} \tag{3.3.15}$$

Now, (3.3.6) follows by linearity from (3.3.5) and (3.3.15) and by the imbedding theorem.

As in §2, define $\tilde{\mathbf{z}}$, $\tilde{p}$, $\tilde{\mathbf{g}}$ and $\tilde{\varphi}$ by $\tilde{\mathbf{z}}(x,y) \stackrel{\text{def}}{=} \mathbf{z}^u(x, y+u(x)), \tilde{p}(x,y) \stackrel{\text{def}}{=} p^u(x, y+u(x)), \tilde{\mathbf{g}}(x,y) \stackrel{\text{def}}{=} \mathbf{g}(x, y+u(x))\sqrt{1+u'^2(x)}, \tilde{\varphi}(x,y) \stackrel{\text{def}}{=} \varphi(x, y+u(x))$. Define operator L by $Lv = \Delta v + v_{yy}(u_x)^2 - 2v_{xy}u_x - v_y u_{xx}$, and $\tilde{\nabla}$ by $\tilde{\nabla}\tilde{p} = \langle \tilde{p}_x - \tilde{p}_y u_x, \tilde{p}_y\rangle$. Then, as before, L is uniformly elliptic.

Let, also, Ξ_u be the map with the image Ω given by the formula $\Xi_u(x,y) = (x, y+u(x))$. Then, $(-\nu\Delta\mathbf{z}^u + \nabla p^u) \circ \Xi_u = -\nu L\tilde{\mathbf{z}} + \tilde{\nabla}\tilde{p}$, and since $|\det D\Xi_u| = 1$ (here $D\Xi_u$ is the gradient matrix of the map Ξ_u so that $|\det D\Xi_u|$ is the Jacobian)

$$\left(-\nu L\tilde{\mathbf{z}} + \tilde{\nabla}\tilde{p}\right)(\tilde{\varphi}) = (-\nu\Delta\mathbf{z}^u + \nabla p^u)(\varphi). \tag{3.3.16}$$

Hence

$$\begin{aligned} \left(-\nu L\tilde{\mathbf{z}} + \tilde{\nabla}\tilde{p}\right)(\tilde{\varphi}) &= \int_{\Gamma_u} \mathbf{g}\cdot\varphi d\sigma \\ &= \int \mathbf{g}(x,u(x))\cdot\varphi(x,u(x))\sqrt{1+u'^2(x)}dx \\ &= \int_{\{y=0\}} \tilde{\mathbf{g}}\cdot\tilde{\varphi}dx \stackrel{\text{def}}{=} \tilde{\xi}(\tilde{\varphi}). \end{aligned} \tag{3.3.17}$$

So,

$$-\nu L\tilde{\mathbf{z}} + \tilde{\nabla}\tilde{p} = \tilde{\xi} \tag{3.3.18}$$

in the sense of distributions. Then from (3.3.18) we get

$$\left(-\nu L\tilde{\mathbf{z}} + \tilde{\nabla}\tilde{p}\right)\left(\delta^1_{-h}\tilde{\varphi}\right) = \tilde{\xi}\left(\delta^1_{-h}\tilde{\varphi}\right). \tag{3.3.19}$$

We have

$$\begin{aligned} \tilde{\xi}\left(\delta^1_{-h}\tilde{\varphi}\right) &= \int_{\{y=0\}} \tilde{\mathbf{g}}\cdot\delta^1_{-h}\tilde{\varphi}dx = -\int_{\{y=0\}} \left(\delta^1_h\tilde{\mathbf{g}}\right)\cdot\tilde{\varphi}dx \\ &\longrightarrow -\int_{\{y=0\}} \tilde{\mathbf{g}}_x\cdot\tilde{\varphi}dx \stackrel{\text{def}}{=} -\tilde{\xi}_x(\tilde{\varphi}), \end{aligned} \tag{3.3.20}$$

as $h \to 0$. We conclude that $(-\nu L\tilde{\mathbf{z}} + \tilde{\nabla}\tilde{p})_x(\tilde{\varphi}) = \tilde{\xi}_x(\tilde{\varphi})$, and hence

$$\begin{aligned} &-\nu L_1\tilde{\mathbf{z}}_x + \tilde{\nabla}\tilde{p}_x \\ &= \tilde{\xi}_x + \nu\left(\tilde{\mathbf{z}}_{yy}2u_xu_{xx} - \tilde{\mathbf{z}}_y u_{xxx}\right) + \left\langle \tilde{p}_y u_{xx}, 0\right\rangle, \end{aligned} \tag{3.3.21}$$

where $L_1 w = \Delta w + (u_x)^2 w_{yy} - 2u_x w_{xy} - 3u_{xx}w_y$. The rest of the proof is similar to the proof of (2.3.8).

Again, in order to claim existence of a minimizer, i.e., existence of a solution of the relaxed problem, one needs some kind of compactness. So let

$$U_b = \{u \in U; \|u\|_{H^3(-1,1)} \leq b\}, \tag{3.3.22}$$

where b is some prescribed positive constant.

PROPOSITION 3.3.1. *Let $\mathbf{g} \in W^{1,q}(\Omega)^2$, for some $q > 2$. Then, there exists an $u^* \in U_b$ such that*

$$\Phi(u^*) = \min_{u \in U_b} \Phi(u). \tag{3.3.23}$$

Proof. The proof is similar to the proof of Proposition 2.3.1.

3.4. Differentiability properties of the variational functional Φ. Our goal is to derive information about the *multivalued* generalized gradient of Φ.

Define the adjoint variable ζ^u, as a solution of the (adjoint) equation

$$\begin{aligned} -\nu\Delta\zeta^u + \nabla q = \eta_u \ \text{ in } \ \mathcal{D}'(\Omega)^2, \ \nabla\cdot\zeta^u = 0 \ \text{ a.e. in } \ \Omega, \\ \zeta^u = 0 \ \text{ on } \ \partial\Omega, \end{aligned} \tag{3.4.1}$$

where $\eta_u \in H^{-1}(\Omega)$ is a vector (signed) measure given by

$$\eta_u(\varphi) = \int_{\Gamma_u} \mathbf{z}^u \cdot \varphi d\sigma. \tag{3.4.2}$$

Obviously, (3.4.1) is the same type of equation as (3.2.2).

In this section, as before, $\mathbf{z}^{u,\text{ext}} = \mathbf{z}^u|_{\Omega_u}$ and $\mathbf{z}^{u,\text{int}} = \mathbf{z}^u|_{\Omega\setminus\Omega_u}$; also, below we shall use the notation $\zeta^{u,\text{ext}} = \zeta^u|_{\Omega_u}$ and $\zeta^{u,\text{int}} = \zeta^u|_{\Omega\setminus\Omega_u}$.

LEMMA 3.4.1. *Let $\mathbf{g} \in W^{2,q}(\Omega)^2$, for some $q > 1$. Then*

$$\zeta^{u,\text{ext}} \in W^{2,q}(\Omega_u)^2, \ \zeta^{u,\text{int}} \in W^{2,q}_{s-\text{loc}}(\Omega\setminus\bar{\Omega}_u)^2. \tag{3.4.3}$$

and the a priori estimates

$$\|\zeta^{u,\text{ext}}\|_{W^{2,q}(\Omega_u)^2} \leq c\left(\|u\|_{H^3(-1,1)}, \|\mathbf{g}\|_{W^{2,q}(\Omega)^2}, \|\mathbf{h}\|_{W^{2-\frac{1}{q},q}(\Omega)^2}\right), \tag{3.4.4}$$

and

$$\begin{aligned} &\|\ \zeta^{u,\text{int}}\|_{W^{2,q}(\Omega\setminus S_\epsilon)} \\ &\leq c\left(\epsilon, \|u\|_{H^3(-1,1)}, \|g\|_{W^{2,q}(\Omega)}, \|\psi\|_{W^{2,q}(\Omega)}\right) \end{aligned} \tag{3.4.5}$$

hold.

Proof. The proof is the same as the proof of Lemma 2.4.1.

We have the next theorem.

THEOREM 3.4.1. *Assume* (3.1.5) *and* (3.1.6). *Then Φ is directionally differentiable at any $u \in U$ such that $u(x) > 0$ for $-1 < x < 1$, and*

$$\begin{aligned} &\Phi'(u;v) \\ &= \int_{-1}^{1} \left(\mathbf{z}^u \cdot (\mathbf{z}^{u,\text{ext}}_y v^+ - \mathbf{z}^{u,\text{int}}_y v^-)\sqrt{1+u'^2} + \frac{1}{2}|\mathbf{z}^u|^2 \frac{u'v'}{\sqrt{1+u'^2}}\right) dx \\ &\quad + \int_{\Gamma_u} \left((\mathbf{g}\cdot\zeta^{u,\text{ext}})_y v^+ - (\mathbf{g}\cdot\zeta^{u,\text{int}})_y v^-\right) d\sigma \\ &\quad + \int_{-1}^{1} \mathbf{g}\cdot\zeta^u \frac{u'v'}{\sqrt{1+u'^2}} dx. \end{aligned} \tag{3.4.6}$$

Moreover, if

$$\mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{int}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{int}})_y \le \mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{ext}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{ext}})_y \quad a.e.\ in\ (-1,1), \tag{3.4.7}$$

then Φ *is subdifferentiable at* u *and*

$$\begin{aligned}
&\partial_* \Phi(u) \\
&= \left[\left(\mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{int}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{int}})_y\right) \sqrt{1+u'^2}, \left(\mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{ext}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{ext}})_y\right) \sqrt{1+u'^2} \right] \\
&\quad - \left(\frac{u'}{\sqrt{1+u'^2}} \left(\frac{1}{2} |\mathbf{z}^u|^2 + \mathbf{g} \cdot \zeta^u \right) \right)'
\end{aligned}$$

$$\stackrel{\mathrm{def}}{=} [l\partial_*\Phi(u), r\partial_*\Phi(u)] \subset L^\infty(-1,1). \tag{3.4.8}$$

On the other hand, if

$$\mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{int}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{int}})_y \ge \mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{ext}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{ext}})_y \quad a.e.\ in\ (-1,1), \tag{3.4.9}$$

then Φ *is superdifferentiable at* u *and*

$$\begin{aligned}
&\partial^* \Phi(u) \\
&= \left[\left(\mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{ext}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{ext}})_y\right) \sqrt{1+u'^2}, \left(\mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{int}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{int}})_y\right) \sqrt{1+u'^2} \right] \\
&\quad - \left(\frac{u'}{\sqrt{1+u'^2}} \left(\frac{1}{2} |\mathbf{z}^u|^2 + \mathbf{g} \cdot \zeta^u \right) \right)'
\end{aligned}$$

$$\stackrel{\mathrm{def}}{=} [l\partial^*\Phi(u), r\partial^*\Phi(u)] \subset L^\infty(-1,1). \tag{3.4.10}$$

Proof. As before, we compute

$$\begin{aligned}
\Phi'(u;v) &= \lim_{\lambda \downarrow 0} \frac{\Phi(u+\lambda v) - \Phi(u)}{\lambda} \\
&= \lim_{\lambda \downarrow 0} \frac{1}{2\lambda} \left(\int_{\Gamma_{u+\lambda v}} |\mathbf{z}^{u+\lambda v}|^2 d\sigma - \int_{\Gamma_u} |\mathbf{z}^u|^2 d\sigma \right) \\
&= \int_{-1}^{1} \left(\mathbf{z}^u \cdot (\mathbf{z}_y^{u,\mathrm{ext}} v^+ - z_y^{u,\mathrm{int}} v^-) \sqrt{1+u'^2} + \frac{1}{2} |\mathbf{z}^u|^2 \frac{u'v'}{\sqrt{1+u'^2}} \right) dx \\
&\quad + \lim_{\lambda \downarrow 0} \frac{1}{2\lambda} \int_{\Gamma_u} \left(|\mathbf{z}^{u+\lambda v}|^2 - |\mathbf{z}^u|^2 \right) d\sigma.
\end{aligned} \tag{3.4.11}$$

LEMMA 3.4.2. *Under previous assumptions on* u *and* v, *and for any* $\alpha < 1$, *the following estimate holds*

$$\|\mathbf{z}^{u+\lambda v} - \mathbf{z}^u\|_{C^0(\bar{\Omega})^2} \le c\lambda^\alpha. \tag{3.4.12}$$

Proof. Let, as before, Ξ_u be the map with the image Ω given by the formula $\Xi_u(x,y) = (x, y + u(x))$. Then $\Xi_u^{-1}(x,y) = (x, y - u(x))$, and (set $A = (x,y)$) $\operatorname{dist}\left(\Xi_{u+\lambda v}^{-1}(A) - \Xi_u^{-1}(A)\right) \le c\lambda$. Now consider $\tilde{\mathbf{z}}^{u+\lambda v}, \tilde{\mathbf{z}}^u, \tilde{p}^{u+\lambda v}, \tilde{p}^u$ defined as $\tilde{\mathbf{z}}^{u+\lambda v} = \mathbf{z}^{u+\lambda v} \circ \Xi_{u+\lambda v}, \tilde{\mathbf{z}}^u = \mathbf{z}^u \circ \Xi_u, \tilde{p}^{u+\lambda v} = p^{u+\lambda v} \circ \Xi_{u+\lambda v}, \tilde{p}^u = p^u \circ \Xi_u$, and operators L_u

and $L_{u+\lambda v}$ defined by

$$\begin{aligned} L_u w &= \Delta w + w_{yy}(u_x)^2 - 2w_{xy}u_x - w_y u_{xx}, \\ L_{u+\lambda v} w &= \Delta w + w_{yy}(u_x+\lambda v_x)^2 - 2w_{xy}(u_x+\lambda v_x) - w_y(u_{xx}+\lambda v_x) \\ &= L_u w + \lambda\left[w_{yy}(2u_x v_x + \lambda v_x^2) - 2w_{xy}v_x - w_y v_{xx}\right], \end{aligned} \tag{3.4.13}$$

and

$$\begin{aligned} \tilde{\nabla}_u w &= \langle w_x - w_y u_x, w_y\rangle, \\ \tilde{\nabla}_{u+\lambda v} w &= \langle w_x - w_y(u+\lambda v)_x, w_y\rangle = \tilde{\nabla}_u w - \langle \lambda v_x, 0\rangle. \end{aligned} \tag{3.4.14}$$

Then $\tilde{\mathbf{z}}^{u+\lambda v} - \tilde{\mathbf{z}}^u$ and corresponding $\tilde{p}^{u+\lambda v} - \tilde{p}^u$ satisfy the equation

$$\begin{aligned} -\nu L_u\left(\tilde{\mathbf{z}}^{u+\lambda v} - \tilde{\mathbf{z}}^u\right) + \tilde{\nabla}_u\left(\tilde{p}^{u+\lambda v} - \tilde{p}^u\right) \\ = \gamma + \langle \lambda v_x, 0\rangle \\ +\nu\lambda\left[\tilde{\mathbf{z}}_{yy}^{u+\lambda v}(2u_x v_x + \lambda v_x^2) - 2\tilde{\mathbf{z}}_{xy}^{u+\lambda v} v_x - \tilde{\mathbf{z}}_y^{u+\lambda v} v_{xx}\right] \end{aligned} \tag{3.4.15}$$

in $\Xi_{u+\lambda v}^{-1}(\Omega) \cap \Xi_u^{-1}(\Omega)$, where

$$\gamma(\varphi) \stackrel{\text{def}}{=} \int_{\{y=0\}} (\mathbf{G}_1 - \mathbf{G}_2)\,\varphi dx \tag{3.4.16}$$

and where

$$\mathbf{G}_1(x,y) \stackrel{\text{def}}{=} \mathbf{g}(x, y+u(x)+\lambda v(x))\sqrt{1+(u'(x)+\lambda v'(x))^2}, \tag{3.4.17}$$

$$\mathbf{G}_2(x,y) \stackrel{\text{def}}{=} \mathbf{g}(x, y+u(x))\sqrt{1+(u'(x))^2}. \tag{3.4.18}$$

Observe that

$$\|\mathbf{G}_1 - \mathbf{G}_2\|_{W^{1,q}\left(\Xi_{u+\lambda v}^{-1}(\Omega)\cap\Xi_u^{-1}(\Omega)\right)^2} \le c\lambda. \tag{3.4.19}$$

Now since $\operatorname{dist}\left(\partial\left(\Xi_{u+\lambda v}^{-1}(\Omega)\right), \partial\left(\Xi_u^{-1}(\Omega)\right)\right) \le c\lambda$ and because of the Hölder continuity of $\mathbf{z}^{u+\lambda v}$ and $\mathbf{z}^u$, we conclude that

$$\|\tilde{\mathbf{z}}^{u+\lambda v} - \tilde{\mathbf{z}}^u\|_{C^0\left(\partial\left(\Xi_{u+\lambda v}^{-1}(\Omega)\cap\Xi_u^{-1}(\Omega)\right)\right)^2} \le c\lambda^\alpha. \tag{3.4.20}$$

Then (3.3.6), (3.4.19), and (3.4.20) imply that

$$\|\tilde{\mathbf{z}}^{u+\lambda v} - \tilde{\mathbf{z}}^u\|_{C^0\left(\overline{\Xi_{u+\lambda v}^{-1}(\Omega)\cap\Xi_u^{-1}(\Omega)}\right)^2} \le c\lambda^\alpha. \tag{3.4.21}$$

Then we have (set $A = (x,y)$)

$$\begin{aligned} &\left|\mathbf{z}^{u+\lambda v}(A) - \mathbf{z}^u(A)\right| \\ &\quad = \left|\tilde{\mathbf{z}}^{u+\lambda v}\left(\Xi_{u+\lambda v}^{-1}(A)\right) - \tilde{\mathbf{z}}^u\left(\Xi_u^{-1}(A)\right)\right| \\ &\quad \le \left|\tilde{\mathbf{z}}^{u+\lambda v}\left(\Xi_{u+\lambda v}^{-1}(A)\right) - \tilde{\mathbf{z}}^u\left(\Xi_{u+\lambda v}^{-1}(A)\right)\right| \\ &\qquad + \left|\tilde{\mathbf{z}}^u\left(\Xi_{u+\lambda v}^{-1}(A)\right) - \tilde{\mathbf{z}}^u\left(\Xi_u^{-1}(A)\right)\right| \\ &\qquad\qquad c\lambda^\alpha + c\lambda^\alpha = c\lambda^\alpha. \end{aligned} \tag{3.4.22}$$

In (3.4.22) we also used the Hölder continuity of $\tilde{\mathbf{z}}^u$. This completes the proof of the lemma.

COROLLARY 3.4.1.

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \int_{\Gamma_u} \left| \mathbf{z}^{u+\lambda v} - \mathbf{z}^u \right|^2 d\sigma = 0. \tag{3.4.23}$$

Now, we can proceed with the proof of the theorem. We compute the last term in (3.4.11).

$$\begin{aligned}
&\lim_{\lambda \downarrow 0} \frac{1}{2\lambda} \int_{\Gamma_u} \left(|\mathbf{z}^{u+\lambda v}|^2 - |\mathbf{z}^u|^2 \right) d\sigma \\
&= \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \int_{\Gamma_u} (\mathbf{z}^{u+\lambda v} - \mathbf{z}^u) \cdot \mathbf{z}^u d\sigma + \lim_{\lambda \downarrow 0} \frac{1}{2\lambda} \int_{\Gamma_u} \left| \mathbf{z}^{u+\lambda v} - \mathbf{z}^u \right|^2 d\sigma \\
&= \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \int_{\Gamma_u} (\mathbf{z}^{u+\lambda v} - \mathbf{z}^u) \cdot \mathbf{z}^u d\sigma = \lim_{\lambda \downarrow 0} \frac{\nu}{\lambda} \int_{\Omega} \nabla \zeta^u : \nabla (\mathbf{z}^{u+\lambda v} - \mathbf{z}^u) \\
&= \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left(\int_{\Gamma_{u+\lambda v}} \mathbf{g} \cdot \zeta^u d\sigma - \int_{\Gamma_u} \mathbf{g} \cdot \zeta^u d\sigma \right) \\
&= \int_{\Gamma_u} \left((\mathbf{g} \cdot \zeta^{u,\mathrm{ext}})_y v^+ - (\mathbf{g} \cdot \zeta^{u,\mathrm{int}})_y v^- \right) d\sigma \\
&\quad + \int_{-1}^{1} \mathbf{g} \cdot \zeta^u \frac{u'v'}{\sqrt{1+u'^2}} dx.
\end{aligned} \tag{3.4.24}$$

Now from (3.4.11) and (3.4.24) we conclude that Φ is directionally differentiable, and that (3.4.6) holds. Furthermore, if (3.4.7) holds, then

$$\begin{aligned}
&\Phi'(u;v) \\
&\quad = \int_{-1}^{1} \left(\mathbf{z}^u \cdot (\mathbf{z}_y^{u,\mathrm{ext}} v^+ - \mathbf{z}_y^{u,\mathrm{int}} v^-) \sqrt{1+u'^2} + \frac{1}{2} |\mathbf{z}^u|^2 \frac{u'v'}{\sqrt{1+u'^2}} \right) dx \\
&\qquad + \int_{\Gamma_u} \left((\mathbf{g} \cdot \zeta^{u,\mathrm{ext}})_y v^+ - (\mathbf{g} \cdot \zeta^{u,\mathrm{int}})_y v^- \right) d\sigma + \int_{-1}^{1} \mathbf{g} \cdot \zeta^u \frac{u'v'}{\sqrt{1+u'^2}} dx \\
&\quad \geq \int_{-1}^{1} \left(\tau - \left(\frac{u'}{\sqrt{1+u'^2}} \left(\frac{1}{2} |\mathbf{z}^u|^2 + \mathbf{g} \cdot \zeta^u \right) \right)' \right) v dx
\end{aligned} \tag{3.4.25}$$

for all

$$\begin{aligned}
\tau \in \Big[&\left(\mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{int}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{int}})_y \right) \sqrt{1+u'^2}, \\
&\left(\mathbf{z}^u \cdot \mathbf{z}_y^{u,\mathrm{ext}} + (\mathbf{g} \cdot \zeta^{u,\mathrm{ext}})_y \right) \sqrt{1+u'^2} \Big].
\end{aligned} \tag{3.4.26}$$

This proves that Φ is subdifferentiable at u and that (3.4.8) holds. Similarly, one can consider superdifferentiability of Φ. So the theorem follows.

Acknowledgment. We thank Eduardo Casas for useful discussions.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] H. W. ALT AND L. A. CAFFARELLI, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 105 (1981), pp. 105–144.
[3] H. W. ALT, L. A. CAFFARELLI, AND A. FRIEDMAN, *A free boundary problem for quasi-linear elliptic equations*, Ann. Scuola Norm. Sup. Pisa, 11 (1984), pp. 1–44.
[4] V. BARBU AND S. STOJANOVIC, *A variational approach to a free boundary problem arising in electrophotography*, Numer. Funct. Anal. Optim., 14 (1993), pp. 1–14.
[5] L. CATTABRIGA, *Su un problema al contorno relativo al sistema di equazioni di Stokes*, Rendiconti Sem. Mat. Univ. Padova, 31 (1961), pp. 308–340.
[6] P. CONSTANTIN AND C. FOIAS, *Navier-Stokes Equations*, The University of Chicago Press, Chicago, IL, 1990.
[7] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.
[8] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, Wiley-Interscience, New York, 1982.
[9] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[10] V. GIRAULT AND P.-A. RAVIART, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[11] B. KAWOHL, *Rearrangements and Convexity of Level Sets in PDE*, LNM #1150, Springer-Verlag, Berlin, 1985.
[12] J. NEČAS, *Les Methodes Directes en Theorie des Equations Elliptiques*, Masson, Paris, 1967.
[13] J. PIPHER, manuscript.
[14] S. STOJANOVIC, *Parallel computations for variational free boundary problem modeling injection of fluid from a slot into a stream*, in Theoretical Aspects of Industrial Design, D.A. Field and V. Komkov, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.
[15] ———, *Nonsmooth Analysis and Shape Optimization in a Flow Problem*, Proceedings of the 31st IEEE Conf. on Decision and Control, pp. 3117–3118, 1992; IMA Preprint Series 1046.
[16] G. M. TROIANIELLO, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 659–674, May 1995

008

ANALYSIS OF A ONE-DIMENSIONAL MODEL FOR COMPRESSIBLE MISCIBLE DISPLACEMENT IN POROUS MEDIA*

YOUCEF AMIRAT† AND MOHAND MOUSSAOUI‡

Abstract. We consider a one-dimensional model problem for the motion of compressible miscible fluids in porous media, without molecular diffusion or dispersion, governed by a nonlinear hyperbolic-parabolic system. We assume the viscosity to be constant. We establish an existence result for nonsmooth data for the concentration. In the case of Dirichlet boundary conditions for the pressure, assuming the data to be smooth, we prove the existence and uniqueness of a smooth solution.

Key words. nonlinear equations, hyperbolic-parabolic system, porous media

AMS subject classifications. 35B40, 76S05, 76T05

1. Introduction.

We are concerned with the single-phase miscible displacement of one compressible fluid by another in a porous medium, under the assumptions that the fluids are miscible in all proportions and that there is no volume change due to the mixing of components (in the sense of Scheidegger [10]). The composition of the mixture is given by the mass concentration. We neglect the molecular diffusion and dispersion and omit the gravitational terms. We only consider the one-dimensional case which can be viewed as modelling experiments in a core sample. The equations corresponding to the description given are as follows (see Chavent and Jaffré [4], Douglas and Roberts [5]):

$$a(x,u)\,\partial_t p + \partial_x q = 0, \tag{1.1}$$

$$\phi(x)\,\partial_t u + q\,\partial_x u + b(x,u)\,\partial_t p = 0, \tag{1.2}$$

$$q = -\frac{k(x)}{\mu(u)}\partial_x p \tag{1.3}$$

for $x \in \Omega$, $t > 0$. Here $\Omega =]0,1[$ represents the core sample, $q(x,t)$ is the single-phase Darcy velocity, $k(x)$ the absolute permeability, $\phi(x)$ the porosity, $\mu(u)$ the viscosity of the mixture, $p(x,t)$ the pressure, $u(x,t)$ the concentration, and

$$a(x,u) = \phi(x)\,(u\,(z_1 - z_2) + z_2), \tag{1.4}$$

$$b(x,u) = \phi(x)\,(z_1 - z_2)\,u\,(1-u), \tag{1.5}$$

where z_i, $i = 1,2$, is the constant compressibility factor for the ith component. In addition to (1.1)–(1.5), we consider the initial and boundary conditions (of mixed type for the pressure):

$$p(0,t) = p_1 \quad \text{(constant)}, \quad q(1,t) = q_1(t), \quad p(x,0) = p_0(x), \tag{1.6}$$

$$u(0,t) = u_1(t), \quad u(x,0) = u_0(x), \tag{1.7}$$

*Received by the editors June 22 1993; accepted for publication (in revised form) November 3, 1993.

†Laboratoire de Mathématiques Appliquées, URA CNRS 1501, Université Blaise Pascal, Clermont-Ferrand-II, 63177 Aubière Cedex, France.

‡Département Mathematiques Informatique et Systèmes, URA CNRS 740, Ecole Centrale de Lyon, 36 avenue Guy de Collongue, 69131 Ecully Cedex, France.

for $x \in \Omega$, $t > 0$. The aim of the present paper is to discuss the existence of a solution (p, u) under reasonable conditions on the data p_1, q_1, p_0, u_0 and u_1. In this study, we assume the viscosity μ to be constant, and for case of exposition, we consider a model problem by taking $\phi \equiv 1$, $\frac{k}{\mu} \equiv 1$, $z_1 = 2$, $z_2 = 1$. The case where μ depends of u is more complicated; we intend to address it in a forthcoming publication.

The paper is organized as follows. Section 2 introduces the notations and hypotheses of the model problem. In §3, we state the existence theorem and describe the method employed. Roughly speaking, our results are as follows. If the data q_1 and p_0 are sufficiently smooth and u_0 and u_1 are functions of finite total variation, then there exists a solution (p, u) where p is a smooth function and u is of finite total variation. This is obtained by the use of Schauder's fixed point theorem. Section 4 is devoted to the proof of the existence theorem. In the final section, we discuss the system (1.1)–(1.5) provided with initial and boundary conditions, of Dirichlet type for the pressure. Assuming the data to be smooth, we then prove that this problem admits a unique smooth solution (p, u).

Note that (1.1)–(1.5) have been considered in Amirat, Hamdache, and Ziani [2], [3] for the related homogenization problems in two cases. For incompressible fluids, since $z_1 = z_2 = 0$, then $a(x, u) = b(x, u) = 0$; for two fluids with the same compressibility factor z, $b(x, u) = 0$ and $a(x, u) = \phi(x)\, z$.

2. The model problem. Let T be a strictly positive real number. We set $\Omega =]0, 1[$ and $Q = \Omega \times]0, T[$. Let a and b denote the functions defined as

$$a(s) = \begin{cases} 1 & \text{if} \quad s \le 0, \\ 1 + s & \text{if} \quad 0 \le s \le 1, \\ 2 & \text{otherwise;} \end{cases} \qquad b(s) = \begin{cases} s(1 - s) & \text{if} \quad 0 \le s \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

We seek two functions u and p defined on Q, solutions of the nonlinear parabolic-hyperbolic system

$$a(u)\, \partial_t p - \partial_x^2 p = 0, \tag{2.2}$$

$$\partial_t u - (\partial_x p)\, \partial_x u + b(u)\, \partial_t p = 0, \tag{2.3}$$

satisfying the boundary and initial conditions

$$p(0, t) = p_1 \quad \text{(constant)}, \quad -\partial_x p(1, t) = q_1(t), \quad t \in]0, T[, \tag{2.4}$$

$$p(x, 0) = p_0(x), \quad x \in \Omega, \tag{2.5}$$

$$u(0, t) = u_1(t), \quad t \in]0, T[, \tag{2.6}$$

$$u(x, 0) = u_0(x), \quad x \in \Omega. \tag{2.7}$$

Observe that we may suppose $p_1 = 0$ by considering $(p - p_1)$ instead of p. We shall do this in the sequel. The following assumptions are made about the boundary and initial conditions:

$$\begin{aligned} &\text{There exists a function} \quad \overline{p} \in L^2(0, T; H^2(\Omega)) \cap H^1(0, T; L^2(\Omega)) \cap L^\infty(0, T; H^1(\Omega)) \\ &\text{such that} \quad \overline{p}(0, t) = 0, \quad -\partial_x \overline{p}(1, t) = q_1(t), \quad t \in]0, T[, \\ &\qquad\qquad \overline{p}(x, 0) = p_0(x), \quad x \in \Omega. \end{aligned} \tag{2.8}$$

$$\begin{aligned} &\text{There is a constant} \quad M > 0 \quad \text{such that} \\ &0 \le -p_0'(x),\ q_1(t) \le M, \quad x \in \Omega,\ t \in]0, T[, \end{aligned} \tag{2.9}$$

$$0 \le u_0(x),\ u_1(t) \le 1, \quad x \in \Omega,\ t \in]0,T[, \tag{2.10}$$

$$u_0 \in BV(\Omega), \quad u_1 \in BV(0,T). \tag{2.11}$$

We make the following comments about these assumptions.

(i) Conditions (2.8) and (2.9) are fulfilled if, for instance,

$$p_0 \in W^{1,\infty}(\Omega), \quad p_0(0) = 0, \quad -p_0' \ge 0, \quad q_1 \in H^{1/2}(]0,T[) \cap C^0([0,T]), \quad q_1 \ge 0.$$

(ii) The optimal assumption which could replace (2.4) would be

$$p(0,t) = p_1(t), \quad -\partial_x p(1,t) = q_1(t), \quad t \in]0,T[, \tag{2.4}'$$

with $p_1 \in H^{3/4}(]0,T[)$ so that for any fixed u in $L^1(Q)$, $0 \le u \le 1$, there exists a solution p of problem (2.2), (2.4)', (2.5) such that $q = -\partial_x p$ satisfies

$$q(x,t) \ge 0 \quad \text{for almost every} \quad (x,t) \in Q.$$

(iii) Since u stands for a concentration, it is quite natural to impose (2.10). Condition (2.11) seems to be a minimal smoothness condition to solve the transport equation (2.3) in a reasonable sense.

(iv) Equation (2.3), subject to the boundary and initial conditions (2.6), (2.7), is a well-posed problem whenever $-\partial_x p(x,t) \ge 0$ almost everywhere in Q. That is what motivates Hypothesis (2.9) on the one hand and the boundary condition (2.4) on the other for p on $x = 0$.

3. The existence result and the method.

DEFINITION 3.1. *A pair (p,u) is said to be a weak solution of problem* (2.2)–(2.7) *if the following assertions hold true*:

(i) $p \in L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega))$, $q \in L^\infty(Q)$,

(ii) $u \in L^\infty(Q)$, $0 \le u \le 1$,

(iii) *p is solution of* (2.2), (2.4), *and* (2.5),

(iv) *u is a weak solution of* (2.3), (2.6), *and* (2.7), *that is, for any φ in $C_0^1([0,1[J\times[0,T[)$,*

$$\int_Q [u(\partial_t\varphi + \partial_x(q\,\varphi)) - b(u)\,\partial_t p\,\varphi]\,dx\,dt = -\int_0^1 u_0(x)\,\varphi(x,0)\,dx - \int_0^T u_1(t)\,q(0,t)\,\varphi(0,t)\,dt.$$

Above $C_0^1([0,1[J \times [0,T[)$ is the space of C^1-differentiable functions having compact support on $[0,1[J \times [0,T[$.

The main result is the following theorem.

THEOREM 3.2. *Under hypotheses* (2.8)–(2.11), *problem* (2.2)–(2.7) *admits a weak solution (p,u) in the sense of Definition* 3.1. *Moreover,*

(i) $\partial_x u \in L^\infty(0,T;M(\Omega))$,

(ii) $\partial_t u \in L^2(0,T;M(\Omega))$, *where $M(\Omega)$ denotes the space of Radon measures on Ω.*

This will be obtained by the use of Schauder's fixed point theorem. Let us introduce the closed convex set in $L^1(Q)$ defined as

$$\mathcal{C} = \{u \in L^1(Q), \quad 0 \leq u \leq 1 \quad \text{a.e. in} \quad Q\}.$$

The idea is then to start from u in $\mathcal{C}$. The pressure equation (2.2) with conditions (2.4) and (2.5) admits a unique solution p. Next, we consider a solution w of the concentration equation

$$\partial_t w + q\,\partial_x w + b(w)\,\partial_t p = 0,$$

where $q = -\partial_x p$, subject to the boundary and initial conditions (2.6) and (2.7). We then set $\mathcal{T}(u) = w$. So we define a mapping $\mathcal{T}$ from $L^1(Q)$ into itself and the problem (2.2)–(2.7) reduces to show the existence of a fixed point of $\mathcal{T}$. The main difficulties are to show that $\mathcal{T}$ is continuous and that $\mathcal{T}(\mathcal{C})$ is relatively compact in $\mathcal{C}$. In fact, we obtain a solution (p,u) of problem (2.2)–(2.7) by the use of a sequence of perturbations $(\mathcal{T}_\eta)_{\eta>0}$ of the mapping $\mathcal{T}$. The concentration equation is replaced by

$$\begin{cases} \partial_t w_\eta + q_\eta\,\partial_x w_\eta + b(w_\eta)\,\partial_t p_\eta = 0, \\ w_\eta(x,0) = u_0^\eta(x), \quad w_\eta(0,t) = u_1^\eta(t), \end{cases}$$

where $q_\eta, p_\eta, u_0^\eta, u_1^\eta$ are suitable regularizations of q, p, u_0 and u_1, respectively. We then set $\mathcal{T}_\eta(u) = w_\eta$. We establish that for any $\eta > 0$, $\mathcal{T}_\eta$ has a fixed point denoted u_η. Finally, a weak solution (p,u) of (2.2)–(2.7) will be obtained as the limit of a subsequence of $(p_\eta, u_\eta)_{\eta>0}$.

Remark 3.3. There is no regularizing effect for the concentration. Indeed, taking $u_0 \equiv 1$, $u_1 \equiv 0$, and q a strictly positive constant, then (p,u) defined by

$$p(x,t) = q\,x, \qquad u(x,t) = \begin{cases} 1 & \text{if} \quad x - q\,t > 0, \\ 0 & \text{if} \quad x - q\,t < 0, \end{cases}$$

is a solution of the corresponding problem (2.2)–(2.7).

4. Proof of Theorem 3.2. The proof is divided into several steps. First, one has the following proposition.

PROPOSITION 4.1. *Let $a = a(x,t)$ be given in $L^\infty(Q)$, such that $1 \leq a(x,t) \leq 2$ for almost every $(x,t) \in Q$. Assume that conditions* (2.8) *and* (2.9) *are fulfilled. Then, there exists a unique function $p \in L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega))$ solution of*

$$\begin{cases} a(x,t)\,\partial_t p - \partial_x^2 p = 0, \quad (x,t) \in Q, \\ p(0,t) = 0, \quad -\partial_x p(1,t) = q_1(t), \quad t \in]0,T[, \\ p(x,0) = p_0(x), \quad xJ \in \Omega. \end{cases} \tag{4.1}$$

Furthermore, one has the estimates

$$||p||_{L^2(0,T;H^2(\Omega))} + ||\partial_t p||_{L^2(Q)} \leq C, \tag{4.2}$$

where C is a constant depending only on the function $\overline{p}$ in (2.8), *and*

$$0 \leq -\partial_x p(x,t) \leq M \quad \textit{for almost every} \quad (x,t) \in Q, \tag{4.3}$$

where M is the constant in (2.9).

Proof. For a given smooth function a (for instance in $C^0(\overline{Q})$), the existence in

$L^2(0,T;H^2(\Omega))\cap H^1(0,T;L^2(\Omega))$ of a solution of (4.1) is well known; see, for example, Ladyzhenskaya, Solonnikov, and Ural'tseva [8]. Therefore, it suffices to derive estimate (4.2) to obtain at the same time the existence of a solution of (4.1) with a in $L^\infty(Q)$. To do so, we put $\tilde{p} = p - \bar{p}$. One easily verifies that

$$a(x,t)\,\partial_t\tilde{p} - \partial_x^2\tilde{p} = f, \quad (x,t) \in Q, \tag{4.4}$$

$$\begin{aligned} &\tilde{p}(0,t) = 0, \quad \partial_x\tilde{p}(1,t) = 0, \quad t \in]0,T[, \\ &\tilde{p}(x,0) = 0, \quad x \in \Omega. \end{aligned} \tag{4.5}$$

Here $f = -(a\,\partial_t\bar{p} - \partial_x^2\bar{p})$. Multiplying (4.4) by $\partial_t\tilde{p}$ and integrating over $Q_t = \Omega\times]0,t[$ with $0 < t < T$, one obtains

$$\begin{aligned} \int_0^t\int_0^1 a(x,s)\,|\partial_t\tilde{p}|^2\,dx\,ds + \frac{1}{2}\int_0^1 |\partial_x\tilde{p}(x,t)|^2\,dx &= \int_0^t\int_0^1 f(x,t)\,\partial_t\tilde{p}\,dx\,ds \\ &\le \frac{1}{2}\int_0^t\int_0^1 |f(x,t)|^2\,dx\,ds + \frac{1}{2}\int_0^t\int_0^1 |\partial_t\tilde{p}|^2\,dx\,ds. \end{aligned}$$

Since $a(x,t) \ge 1$, one deduces

$$||\partial_t\tilde{p}||^2_{L^2(Q)} + ||\partial_x\tilde{p}||^2_{L^\infty(0,T;L^2(\Omega))} \le 2\,C\,||\bar{p}||^2_W,$$

where $W = L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega)) \cap L^\infty(0,T;H^1(\Omega))$. Hence (4.2) and the existence of a solution of (4.1) follow. The uniqueness is obvious. To prove (4.3), we note that $q = -\partial_x p$ is formally a solution of

$$\begin{cases} \partial_t q - \partial_x\left(\dfrac{1}{a(x,t)}\partial_x q\right) = 0, \quad (x,t) \in Q, \\ \partial_x q(0,t) = 0, \quad q(1,t) = q_1(t), \quad t \in]0,T[, \\ q(x,0) = -p_0'(x), \quad x \in \Omega. \end{cases} \tag{4.6}$$

Since $q \in L^2(0,T;H^1(\Omega))$ and $\partial_t q \in L^2(0,T;H^{-1}(\Omega))$, then $q \in C^0([0,T];H^{-1}(\Omega))$. Therefore the initial condition in (4.6) for q is meaningful. A reflexion procedure with respect to $x = 0$ allows the transformation of (4.6) to a mixed Cauchy–Dirichlet problem over $]-1,1[\times]0,T[$. Then the maximum principle gives (4.3). This completes the proof.

Now let us define the mappings $\mathcal{T}_\eta$ and the approximate fixed points u_η. With any fixed u in $\mathcal{C}$, we associate the solution p of

$$\begin{cases} a(u)\,\partial_t p - \partial_x^2 p = 0, \quad (x,t) \in Q, \\ p(0,t) = 0, \quad -\partial_x p(1,t) = q_1(t), \quad t \in]0,T[, \\ p(x,0) = p_0(x), \quad x \in \Omega. \end{cases} \tag{4.7}$$

Setting $q = -\partial_x p$, p and q then satisfy estimates (4.2) and (4.3). Let us introduce a regularization q_η of q, such that, for any $\eta > 0$, small enough,

(i) $\eta \le q_\eta(x,t) \le M + \eta$, for every $(x,t) \in Q$,

(ii) q_η converges, as η tends to 0, to q strongly in $L^2(Q)$ and weakly in $L^2(0,T;H^1(\Omega))$,

(iii) $|\partial_x(q_\eta)| \le \frac{C}{\eta}$, where C is a constant independent on η.

Such a regularization exists and may be constructed as follows. First one extends q by even symetrization with respect to $x = 0, x = 1, t = 0, t = T$ into a function q_2

defined on $\tilde{Q} =]-1, 2[J\times]-T, 2T[$. Then we choose a cutoff function θ in $\mathcal{D}(\tilde{Q})$ such that $\theta \equiv 1$ on Q and $0 \leq \theta \leq 1$. Then we set

$$\tilde{q} = \begin{cases} \theta q_2 & \text{in} \quad \tilde{Q}, \\ 0 & \text{outside.} \end{cases}$$

One easily verifies that $0 \leq \tilde{q} \leq M$ and $\tilde{q}$ is bounded in $L^2(I\!R_t, H^1(I\!R))$. Next we consider a regularizing kernel ρ with support in the unit ball satisfying $\rho \in \mathcal{D}(I\!R^2)$, $\rho \geq 0$ and $\int_{I\!R^2} \rho(x,t)\,dx\,dt = 1$. Then we set $\rho_\eta(x,t) = \frac{1}{\eta^2}\,\rho(\frac{x}{\eta}, \frac{t}{\eta})$. It is easily seen that the function

$$q_\eta(x,t) = \eta + \int_{I\!R^2} \rho_\eta(x-\xi, t-\tau)\,\tilde{q}(\xi,\tau)\,d\xi\,d\tau \quad \text{for} \quad (x,t) \in Q$$

satisfies (i), (ii), and (iii).

By using similar techniques of regularisation, we define a $C^2(\overline{Q})$ function p_η converging to p in $W = L^2(0,T; H^2(\Omega)) \cap H^1(0,T; L^2(\Omega)) \cap L^\infty(0,T; H^1(\Omega))$, satisfying the following properties:

(i) $p_\eta(0,t) = 0, \quad 0 < t < T,$

(ii) $||p_\eta||_W \leq K\,||p||_W$, K independent of η,

(iii) $||\partial_x p_\eta||_{L^\infty(Q)} \leq M'$ with M' independent of η,

(iv) $||p_\eta||_{C^2(\overline{Q})} \leq K(\eta)\,||p||_W$.

In the sequel M is set in place of $\max(M, M')$.

The regularisation of the data u_0 and u_1 is done through the following lemma.

LEMMA 4.2. *Let $v_0 \in BV(\Omega)$ and $v_1 \in BV(0,T)$ with $0 \leq v_0 \leq 1$, $0 \leq v_1 \leq 1$. Then for any η, $0 < \eta < \frac{1}{2}$, there exist two sequences $(v_0^\eta) \subset C^1(\overline{\Omega})$, $(v_1^\eta) \subset C^1([0,T])$ satisfying*

(i) $v_0^\eta(0) = v_1^\eta(0)$,

(ii) $(v_0^\eta)'(0) = (v_1^\eta)'(0) = 0$,

(iii) $0 \leq v_0^\eta \leq 1, \quad 0 \leq v_1^\eta \leq 1$,

(iv) $||v_0^\eta||_{W^{1,1}(\Omega)} + ||v_1^\eta||_{W^{1,1}(]0,T[)} \leq C\,(||v_0||_{BV(\Omega)} + ||v_1||_{BV(0,T)})$, *where C is a constant independent of η,*

(v) (v_0^η, v_1^η) *converges to (v_0, v_1) in $L^1(\Omega) \times L^1(0,T)$ as η tends to zero.*

Proof. The proof is a consequence of the following.

Let f in $BV(\Omega)$, the latter equipped with the norm $||f||_{BV(\Omega)} = ||f||_{L^\infty(\Omega)} + TV(f)$ (or, equivalently, $||f||_{BV(\Omega)} = ||f||_{L^1(\Omega)} + TV(f)$). Then there exists a sequence (g_η) in $C^1(\overline{\Omega})$ and a constant C independent of η such that

(a) $g_\eta(0) = g_\eta'(0) = 0$,

(b) $||g_\eta||_{W^{1,1}(\Omega)} \leq C||f||_{BV(\Omega)}$,

(c) (g_η) converges to f in $L^1(\Omega)$ as η tends to zero; $0 \leq g_\eta \leq 1$ if $0 \leq f \leq 1$.

First we introduce the sequence (f_η) defined by

$$f_\eta(x) = \begin{cases} f(x) & \text{if} \quad 2\eta \leq x < 1, \\ 0 & \text{if} \quad 0 < x < 2J\eta. \end{cases}$$

Then it is clear that

(1) $||f_\eta||_{L^\infty(\Omega)} \leq ||f||_{L^\infty(\Omega)}, \quad 0 \leq f_\eta \leq 1$ if $0 \leq f \leq 1$,

(2) $TV(f_\eta) \leq C\,||f||_{BV(\Omega)}$, with a constant C independent of η,

(3) $f_\eta \to f$ in $L^p(\Omega)$, for any $p, 1 \leq p < \infty$.

Next we consider the sequence $(\tilde{g}_\eta)$ defined on $\mathbb{R}$ by the convolution

$$\tilde{g}_\eta(x) = \frac{1}{\eta}\int_{-\infty}^{+\infty} \rho\left(\frac{x-y}{\eta}\right) \tilde{f}_\eta(y)\, dy,$$

where $\tilde{f}_\eta$ denotes the extension of f_η by zero outside Ω and ρ is a nonnegative regularizing kernel, $\rho \in \mathcal{D}(]-1,1[)$, $\int_{-\infty}^{+\infty} \rho(y)\, dy = 1$. Then one easily verifies that the sequence (g_η) obtained by restricting each $\tilde{g}_\eta$ to Ω satisfies (a), (b), and (c). Now we define w_η as the solution of

$$\begin{cases} \partial_t w_\eta + q_\eta\, \partial_x w_\eta + b(w_\eta)\, \partial_t p_\eta = 0, & (x,t) \in Q, \\ w_\eta(0,t) = u_1^\eta(t), \quad w_\eta(x,0) = u_0^\eta(x), & t \in]0,T[,\ x \in \Omega, \end{cases} \tag{4.8}$$

and set $\mathcal{T}_\eta(u) = w_\eta$.

PROPOSITION 4.3. *For any $\eta > 0$, the mapping $\mathcal{T}_\eta$ is well defined and has a fixed point in $\mathcal{C}$.*

The proof relies on a priori estimates. We shall note those which are independent of the parameter η.

LEMMA 4.4. *For any u in $\mathcal{C}$, $\mathcal{T}_\eta(u) = w_\eta$ is well defined and satisfies the following estimates*:

$$0 \le w_\eta(x,t) \le 1, \qquad \text{for every} \quad (x,t) \in Q, \tag{4.9}$$

$$||w_\eta||_{L^\infty(0,T;W^{1,1}(\Omega))} + ||\partial_t(w_\eta)||_{L^2(0,T;L^1(\Omega))} \le C_1, \tag{4.10}$$

$$||w_\eta||_{L^\infty(0,T;H^1(\Omega))} + ||\partial_t(w_\eta)||_{L^2(Q)} \le C_2(\eta), \tag{4.11}$$

where C_1 and $C_2(\eta)$ are constants which depend on the data, C_1 independent of η.

Remark 4.5. In Lemma 4.4, the compatibility conditions necessary for the existence of a smooth solution w_η are

$$\begin{aligned} &u_0^\eta \in C^1(\overline{\Omega}), \quad u_1^\eta \in C^1([0,T]), \\ &u_0^\eta(0) = u_1^\eta(0), \quad (u_1^\eta)'(0) + q_\eta(0,0)\,(u_0^\eta)'(0) = 0. \end{aligned} \tag{4.12}$$

Clearly, from Lemma 4.2, these conditions hold. Also notice that, from properties (i) and (iii) of p_η, we have

$$-M \le p_\eta(x,t) \le M. \tag{4.13}$$

Proof of Lemma 4.4. Let $u \in \mathcal{C}$. From Proposition 4.1, problem (4.7) has a unique solution p. We consider now the problem (4.8) where, for convenience we drop the subscript η except for q_η, which will be denoted $\overline{q}$ throughout the sequel; that is,

$$\partial_t w + \overline{q}\, \partial_x w + \partial_t p\, b(w) = 0, \quad (x,t) \in Q, \tag{4.14}$$

$$w(0,t) = u_1(t), \quad t \in]0,T[, \quad w(x,0) = u_0(x), J \quad x \in \Omega. \tag{4.15}$$

Equation (4.14) may be reduced (at least formally) to a linear equation by introducing an antiderivative of $\frac{1}{b(s)}$, namely $G(s) = \ln \frac{s}{(1-s)}$ defined on $]0,1[$. However, the boundary value problem is well posed only if u_0 and u_1 do not assume the values 0 or 1. We introduce a perturbation of (4.14), namely,

$$\partial_t w^\varepsilon + \overline{q}\, \partial_x w^\varepsilon + (b(w^\varepsilon) + \varepsilon(1+\varepsilon))\partial_t p = 0, \tag{4.16}$$

where $\varepsilon > 0$. Setting

$$G^\varepsilon(s) = \frac{1}{1+2\varepsilon} \ln\left(\frac{\varepsilon + s}{\varepsilon + 1 - s}\right), \quad s \in [0,1],$$

it is readily verified that $G^\varepsilon(w^\varepsilon)$ is solution of

$$\partial_t(G^\varepsilon(w^\varepsilon)) + \overline{q}\,\partial_x(G^\varepsilon(w^\varepsilon)) + \partial_t p = 0.$$

Following an idea used by Kazhikhov and Shelukin [6] and Serre [11], we introduce the auxiliary function

$$h^\varepsilon = G^\varepsilon(w^\varepsilon) + p.$$

Thus, h^ε is a solution of

$$\text{(4.17)} \qquad \begin{cases} \partial_t h^\varepsilon + \overline{q}\,\partial_x h^\varepsilon = -q\,\overline{q}, \quad (x,t) \in Q, \\ h^\varepsilon(0,t) \equiv h_1^\varepsilon(t) = \dfrac{1}{1+2\varepsilon} \ln\left(\dfrac{\varepsilon + u_1(t)}{\varepsilon + 1 - u_1(t)}\right), \quad t \in]0,T[, \\ h^\varepsilon(x,0) \equiv h_0^\varepsilon(x) = \dfrac{1}{1+2\varepsilon} \ln\left(\dfrac{\varepsilon + u_0(x)}{\varepsilon + 1 - u_0(x)}\right) + p_0(x), \quad x \in \Omega. \end{cases}$$

According to (4.12) and (4.13), we have

$$h_0^\varepsilon \in C^1(\overline{\Omega}), \quad h_1^\varepsilon \in C^1([0,T]),$$
$$h_0^\varepsilon(0) = h_1^\varepsilon(0), \quad h_1^{\varepsilon\prime}(0) + \overline{q}(0,0)\,h_0^{\varepsilon\prime}(0) = -q(0,0)\,\overline{q}(0,0).$$

Furthermore, q and $\overline{q}$ belong to $C^1(\overline{Q})$ and $\overline{q}$ is positive. Then, problem (4.17) has a unique solution in $C^1(\overline{Q})$. Therefore, the function

$$w^\varepsilon = \frac{(1+\varepsilon)\,e^{(1+2\varepsilon)(h^\varepsilon - p)} - \varepsilon}{1 + e^{(1+2\varepsilon)(h^\varepsilon - p)}}$$

is a solution of (4.16). It remains to prove that $(w^\varepsilon)_{\varepsilon>0}$ converges, as ε tends to 0, to a solution of (4.14) and (4.15). To this end, we introduce the function

$$v^\varepsilon = \frac{e^{(1+2\varepsilon)h^\varepsilon}}{1 + e^{(1+2\varepsilon)h^\varepsilon}}.$$

A straightforward calculation shows that v^ε is a solution of

$$\begin{cases} \partial_t v_\varepsilon + \overline{q}\,\partial_x v_\varepsilon = -(1+2\varepsilon)\,v^\varepsilon(1-v^\varepsilon)\,q\,\overline{q}, \quad (x,t) \in Q, \\ v^\varepsilon(0,t) = \dfrac{\varepsilon + u_1(t)}{1+2\varepsilon}, \quad t \in]0,T[, \\ v_\varepsilon(x,0) = \dfrac{(\varepsilon + u_0(x))\,e^{(1+2\varepsilon)\,p_0(x)}}{(1+\varepsilon - u_0(x)) + (\varepsilon + u_0(x))\,e^{(1+2\varepsilon)\,p_0(x)}}, \quad x \in \Omega. \end{cases}$$

Then, one easily verifies that, as ε tends to 0, $(v^\varepsilon)_{\varepsilon>0}$ converges in $H^1(Q)$ weak and in $L^2(Q)$ strong to a v solution of

$$\text{(4.18)} \qquad \partial_t v + \overline{q}\,\partial_x v = -v(1-v)\,q\,\overline{q}, \quad (x,t) \in Q,$$
$$v(0,t) = u_1(t), \quad t \in]0,T[,$$

$$\text{(4.19)} \qquad v(x,0) = \frac{u_0(x)\,e^{p_0(x)}}{(1-u_0(x)) + u_0(x)\,e^{p_0(x)}}, \quad x \in \Omega,$$

satisfying, in addition, $0 \le v(x,t) \le 1$ for $(x,t) \in Q$. According to (4.12) and (4.13), v belongs to $C^1(\overline{Q})$. On the other hand, w^ε may be written in the form

$$w^\varepsilon = \frac{(1+\varepsilon)\, e^{-(1+2\varepsilon)\, p}\, v^\varepsilon - \varepsilon(1-v^\varepsilon)}{(1-v^\varepsilon) + v^\varepsilon\, e^{-(1+2\varepsilon)p}}.$$

It follows that, as ε tends to 0, $(w^\varepsilon)_{\varepsilon>0}$ converges in $H^1(Q)$ weak to w given by

$$w = \frac{e^{-p}\, v}{(1-v) + v\, e^{-p}}. \tag{4.20}$$

It is easily verified that w is a solution of problem (4.14) and (4.15) and satisfies (4.9). Let us now prove estimates (4.10) and (4.11). They will be a consequence of the following lemma.

LEMMA 4.6. *Let v be the solution of Problem* (4.18) *and* (4.19). *Then*

$$||v||_{L^\infty(0,T;W^{1,1}(\Omega))} + ||v||_{W^{1,\infty}(0,T;L^1(\Omega))} \le C_3, \tag{4.21}$$

$$||v||_{L^\infty(0,T;H^1(\Omega))} + ||v||_{W^{1,\infty}(0,T;L^2(\Omega))} \le C_4(\eta), \tag{4.22}$$

where C_3 and $C_4(\eta)$ are constants, C_3 independent of η.

Indeed, admitting this lemma for a moment, from the inequalities $-M \le p \le 0$ it follows that

$$1 \le (1-v) + v\, e^{-p} \le e^M.$$

Then, using the relation

$$w = \frac{e^{-p}\, v}{(1-v) + v\, e^{-p}}$$

together with (4.2) and (4.3), the estimates (4.10) and (4.11) follow immediately.

Proof of Lemma 4.6. From (4.18) and (4.19) it follows that

$$\begin{aligned} |\overline{q}(0,t)\, \partial_x v(0,t)| &= |-u_1(t)(1-u_1(t))\, q(0,t)\, \overline{q}(0,t) - \partial_t u_1(t)| \\ &\le \frac{1}{4} M(M+\eta) + |\partial_t u_1(t)| \le \frac{1}{2} M^2 + |\partial_t u_1(t)| \qquad \text{for} \quad \eta \le M \end{aligned}$$

for almost every $t \in]0,T[$, and thus

$$\int_0^T |\overline{q}(0,t)\, \partial_x v(0,t)|\, dt \le \frac{1}{2} M^2\, T + \int_0^T |\partial_t u_1(t)|\, dt. \tag{4.23}$$

Now we differentiate (4.18) with respect to x. We get

$$\partial_t(\partial_x v) + \partial_x(\overline{q}\, \partial_x v) = -v(1-v)\, (q\, \partial_x \overline{q} + \overline{q}\, \partial_x q) + q\, \overline{q}\, (2v-1)\, \partial_x v. \tag{4.24}$$

Multiplying (4.24) by $\operatorname{sign}(\partial_x v)$ and integrating in x yields

$$\begin{aligned} \frac{d}{dt} \int_0^1 |\partial_x v|\, dx &+ \int_0^1 \partial_x(\overline{q}\, |\partial_x v|)\, dx \\ &= \int_0^1 v(v-1)\, (q\, \partial_x \overline{q} + \overline{q}\, \partial_x q) \operatorname{sign}(\partial_x v)\, dx + \int_0^1 q\, \overline{q}\, (2v-1)\, |\partial_x v|\, dx. \end{aligned}$$

Hence, by integration in time from 0 to t, $0 < t < T$, we obtain

$$\int_0^1 |\partial_x v(x,t)|\,dx + \int_0^t \overline{q}(1,s)\,|\partial_x v(1,s)|\,ds \le \int_0^1 |\partial_x v(x,0)|\,dx + \int_0^t \overline{q}(0,s)\,|\partial_x v(0,s)|\,ds + \frac{M}{2}\int_0^t\int_0^1 (|\partial_x q| + |\partial_x \overline{q}|)\,dx\,ds + 2\,M^2 \int_0^t\int_0^1 |\partial_x v|\,dx\,ds.$$

Using (4.23), Proposition 4.1 and the definition of $\overline{q}$, we obtain

$$\int_0^1 |\partial_x v(x,t)|\,dx \le K_1 + 2\,M^2 \int_0^t\int_0^1 |\partial_x v|\,dx\,ds$$

for any $t > 0$, $t \le T$, where K_1 is a constant independent of η. Thus, in view of Gronwall's lemma, $\partial_x v$ is bounded in $L^\infty(0,T;L^1(\Omega))$ by a constant depending on the data but which does not depend on η. That $\partial_t v$ is bounded in the same way follows readily from (4.18). Hence (4.21) follows. To prove (4.22), we multiply (4.24) by $\partial_x v$. We get

$$\frac{1}{2}\frac{d}{dt}|\partial_x v|^2 + \partial_x(\overline{q}\,\partial_x v)\,\partial_x v = \partial_x(v(v-1)q\,\overline{q})\,\partial_x v. \tag{4.25}$$

Then, by integration over $Q_t =]0,1[\times]0,t[$, with $0 < t \le T$, we obtain

$$\frac{1}{2}\int_0^1 |\partial_x v(x,t)|^2\,dx + \frac{1}{2}\int_0^t \overline{q}(1,s)|\partial_x v(1,s)|^2\,ds = \frac{1}{2}\int_0^1 |\partial_x v(x,0)|^2\,dx + \frac{1}{2}\int_0^t \overline{q}(0,s)|\partial_x v(0,s)|^2\,ds - \frac{1}{2}\int_0^t\int_0^1 \partial_x\overline{q}(x,s)|\partial_x v(x,s)|^2\,dx\,ds + \int_0^t\int_0^1 (\partial_x(v(v-1)q\overline{q})\,\partial_x v)(x,s)\,dx\,ds.$$

The first term on the right side is majorised in terms of $H^1(\Omega)$ norm of the data p_0 and u_0. For the second, we again use (4.18). We have

$$\begin{aligned}\overline{q}(0,s)\,|\partial_x v(0,s)|^2 &= -u_1(s)(1-u_1(s))\,q(0,s)\,\overline{q}(0,s)\,\partial_x v(0,s) - \partial_t u_1(s)\,\partial_x v(0,s)\\ &\le \frac{1}{4}\,M\sqrt{M+\eta}\,\sqrt{\overline{q}(0,s)}\,|\partial_x v(0,s)| + |\partial_t u_1(s)|\,\frac{1}{\sqrt{\eta}}\cdot\sqrt{\eta}|\partial_x v(0,s)|\\ &\le \frac{M^3}{8} + \frac{1}{4}\,\overline{q}(0,s)\,|\partial_x v(0,s)|^2 + \frac{1}{\eta}\,|\partial_t u_1(s)|^2 + \frac{\eta}{4}\,|\partial_x v(0,s)|^2\\ &\le \frac{M^3}{8} + \frac{1}{2}\,q(0,s)\,|\partial_x v(0,s)|^2 + \frac{1}{\eta}\,|\partial_t u_1(s)|^2 \quad \text{since} \quad \overline{q}(0,s) \ge \eta.\end{aligned}$$

Hence

$$\frac{1}{2}\int_0^t \overline{q}(0,s)\,|\partial_x v(0,s)|^2\,ds \le \frac{M^3}{8}\,t + \frac{1}{\eta}\int_0^t |\partial_t u_1(s)|^2\,ds.$$

For the third term, according to the property (iii) of $\overline{q}$, one has

$$\int_0^t\int_0^1 (\partial_x\overline{q}\,|\partial_x v|^2)(x,s)\,dx\,ds \le \frac{C}{\eta}\int_0^t\int_0^1 |J\partial_x v(x,s)|^2\,dx\,ds.$$

For the last term, one has

$$I = \int_0^t \int_0^1 (2v - 1)\, q\, \bar{q} |\partial_x v|^2 \, dx\, ds + \int_0^t \int_0^1 v(v-1)(q\, \partial_x \bar{q} + \bar{q}\, \partial_x q)\, \partial_x v \, dx\, ds,$$

where

$$I = \int_0^t \int_0^1 \partial_x (v(v-1) q\, \bar{q})\, \partial_x v \, dx\, ds.$$

Hence,

$$I \leq (2M^2 + 1) \int_0^t \int_0^1 |\partial_x v|^2 \, dx\, ds + \frac{M^2}{4} \int_0^t \int_0^1 (|\partial_x \bar{q}|^2 + |\partial_x q|^2)\, dx\, ds.$$

Finally,

$$\frac{1}{2} \int_0^1 |\partial_x v|^2(x,t)\, dx + \frac{1}{2} \int_0^1 \bar{q}(1,s)\, |\partial_x v(1,s)|^2 \, ds \leq K_2 + \frac{K_3}{\eta} + \left(K_4 + \frac{C}{\eta}\right) \int_0^t \int_0^1 |\partial_x v|^2 \, dx\, ds,$$

where K_2, K_3, and K_4 are constants independent of η. Consequently, $\partial_x v$ is bounded in $L^\infty(0,T; L^2(\Omega))$ by a constant which depends on the data, but also on the parameter η. Thanks to (4.18), the same holds true for $\partial_t v$. Lemma 4.6 is proved.

Clearly, by (4.9), $\mathcal{T}_\eta(\mathcal{C})$ is contained in $\mathcal{C}$ and, according to Lemmas 4.2 and 4.4, $\mathcal{T}_\eta(\mathcal{C})$ is relatively compact in $\mathcal{C}$ since by (4.11) it is contained in a bounded set of $H^1(Q)$. To prove the statement of Proposition 4.3, it thus remains to show the following lemma.

LEMMA 4.7. *For any $\eta > 0$, the mapping $\mathcal{T}_\eta$ from $\mathcal{C}$ into itself is continuous for the $L^1(Q)$ norm.*

Proof. Let (u^n) be an arbitrary sequence which converges, as n tends to ∞, to a function u in $L^1(Q)$. In fact, (u^n) converges, as n tends to ∞, to u in $L^r(Q)$, for any r, $1 \leq r < \infty$. The associated sequence (p^n) with (u^n) is bounded in the space $L^2(0,T; H^2(\Omega)) \cap H^1(0,T; L^2(\Omega))$. Then there is a subsequence, still denoted by (p^n), which is weakly convergent. It is easily seen that the limit p of (p^n) is a solution of

$$\begin{cases} a(u)\partial_t p - \partial_x^2 p = 0, & (x,t) \in Q, \\ p(0,t) = 0, \quad -\partial_x p(1,t) = q_1(t), & t \in]0,T[, \\ p(x,0) = p_0(x), & x \in \Omega. \end{cases}$$

Since this problem has a unique solution, the whole sequence (p^n) converges to p. In addition, the sequence $q^n = -\partial_x p^n$ converges, as n tends to ∞, to $q = -\partial_x p$, strongly in $L^2(Q)$. Indeed,

$$(q^n) \quad \text{is bounded in} \quad L^2(0,T; H^1(\Omega)),$$
$$(\partial_t q^n) \quad \text{is bounded in} \quad L^2(0,T; H^{-1}(\Omega)).$$

Consequently, by Aubin's theorem (see Lions [9]), $\{q^n\}$ is compactly imbedded in $L^2(Q)$. Therefore, (q^n) and $(\bar{q}^n)$ converge, respectively, as n tends to ∞, to q and $\bar{q}$ in $L^r(Q)$, for any r, $1 \leq r < \infty$. Let us now consider w and w^n the respective solutions

of problems (4.14) and (4.15) with respective coefficients $\overline{q}$ and $\overline{q}^n$. We also introduce the auxiliary functions v and v^n by the definition

$$v = \frac{e^p\, w}{(1-w) + w\, e^p}, \quad v^n = \frac{e^{p^n}\, w^n}{(1-w^n) + w^n\, e^{p^n}}.$$

Clearly, (v^n) is bounded in $H^1(Q)$ and v is in $H^1(Q)$ and they are solutions of the corresponding equations (4.18), with respective coefficients $(q, \overline{q})$ and $(q^n, \overline{q}^n)$, with the same boundary and initial conditions (4.19). Then, the difference $z^n = v - v^n$ satisfies

$$\begin{aligned}\partial_t z^n + \overline{q}^n\, \partial_x z^n =& (\overline{q}^n - \overline{q})\, \partial_x v + \overline{q}^n\, q^n\, (v + v^n - 1)\, z^n \\ &+ v(v-1)\, (q\,\overline{q} - q^n\, \overline{q}^n).\end{aligned}$$

Multiplying by $\operatorname{sign}(z^n)$ and integrating in x it follows

$$\begin{aligned}\frac{d}{dt}\int_0^1 |z^n|\, dx + \overline{q}^n(1,t)\, |z^n(1,t)| &\le \int_0^1 \partial_x(\overline{q}^n)\, |z^n|\, dx + \int_0^1 |\overline{q}^n - \overline{q}|\, |\partial_x v|\, dx \\ &+ M(M+\eta)\int_0^1 |z^n|\, dx + \frac{1}{4}\int_0^1 |q\,\overline{q} - q^n\,\overline{q}^n|\, dx \\ &\le \left(2\,M^2 + \frac{C}{\eta}\right)\int_0^1 |z^n|\, dx + \int_0^1 |\overline{q}^n - \overline{q}|\, |\partial_x v|\, dx \\ &+ \frac{M}{4}\int_0^1 |\overline{q} - \overline{q}^n|\, dx + \frac{M+\eta}{4}\int_0^1 |q - q^n|\, dx.\end{aligned}$$

Integrating in time, we obtain

$$\begin{aligned}\int_0^1 |z^n(x,t)|\, dx \le \left(2\,M^2 + \frac{C}{\eta}\right)\int_0^t\int_0^1 |z^n(x,s)|\, dx\, ds + \frac{C}{\sqrt{\eta}}\left(\int_0^t\int_0^1 |\overline{q}^n - \overline{q}|^2\, dx\, ds\right)^{1/2} \\ + \frac{M}{2}\int_0^t\int_0^1 (|\overline{q} - \overline{q}^n| + |q - q^n|)\, dx\, ds.\end{aligned}$$

For any $\varepsilon > 0$, we can find an integer N such that $n > N$ implies

$$\int_0^1 |Jz^n(x,t)J|\, dx \le K_5(\eta)\int_0^t\int_0^1 |Jz^n(x,s)J|\, dx\, ds + K_6(\eta)\,\varepsilon,$$

where $K_5(\eta)$ and $K_6(\eta)$ are constants. So, putting $y^n(t) = \int_0^1 |z^n(x,t)|\, dx$ we obtain, with Gronwall's lemma,

$$|y^n(t)| \le C(\eta)\,\varepsilon.$$

This means that v^n converges, as n tends to ∞, to v in $L^\infty(0,T; L^1(\Omega))$. Using the fact that (p^n) is uniformly bounded and weakly convergent to p, together with the relations linking w and v, one readily deduces that (w^n) converges, as n tends to ∞, to w in $L^r(Q)$, for any r, $1 \le r < \infty$. Lemma 4.7 is proved. Hence Proposition 4.3 follows.

We are now in a position to prove Theorem 3.2.

Proof of Theorem 3.2. Take a real sequence (η_n) which tends to 0 as n tends to ∞. For any n, there is a pair (p^n, u^n) solution of

$$a(u^n)\, \partial_t p^n - \partial_x^2 p^n = 0, \quad (x,t) \in Q, \tag{4.26}$$

$$\partial_t u^n + \overline{q}^n\, \partial_x u^n + \partial_t p^n\, b(u^n) = 0, \quad (x,t) \in Q, \tag{4.27}$$

satisfying the boundary and initial conditions (2.4)–(2.7), after a regularization in order to verify (4.12) and (4.13). Since (u^n) is bounded in the space $L^\infty(0,T;W^{1,1}(\Omega))\cap H^1(0,T;L^1(\Omega))$ (see Remark 4.5), there is a subsequence, still denoted (u^n) which converges to a function u in $L^1(Q)$ and therefore also in $L^r(Q)$ for any r, $1 \le r < \infty$. The sequence (p^n) is bounded in $L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega))$. Then there is a subsequence, still denoted (p^n), which converges, as n tends to ∞, weakly in $L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega))$ to a function p. By a compacity argument, $q^n = -\partial_x p^n$ converges, as n tends to ∞, to $q = -\partial_x p$ in $L^2(Q)$ for any r, $1 \le r < \infty$. Therefore, $\overline{q}^n$ converges, as n tends to ∞, to q in $L^r(Q)$ for any r, $1 \le r < \infty$, and also in $L^2(0,T;H^1(\Omega))$ weak. Trivially, $a(u^n)$ (resp., $b(u^n)$) converges, as n tends to ∞, to $a(u)$ (resp., $b(u)$) in $L^r(Q)$, for any r, $1 \le r < \infty$. Then, we can pass to the limit in (4.26). We obtain

$$\begin{cases} a(u)\partial_t p - \partial_x^2 p = 0, & (x,t) \in Q, \\ p(0,t) = 0, \quad q(1,t) = q_1(t), & t \in]0,T[, \\ p(x,0) = p_0(x), & x \in \Omega. \end{cases}$$

On the other hand, for any n and any φ in $C_0^1([0,1[J \times [0,T[)$,

$$\int_0^T \int_0^1 (u^n\, \partial_t \varphi + u^n\, \partial_x(\overline{q}^n\, \varphi) - b(u^n)\, \partial_t p^n\, \varphi)\, dx\, dt$$
$$= -\int_0^1 u_0(x)\, \varphi(x,0)\, dx - \int_0^T u_1(t)\, \overline{q}^n\, \varphi(0,t)\, dt,$$

and passing to the limit in each term is meaningful. Note in particular that the convergence of $\overline{q}^n$ to q in $L^2(0,T;H^1(\Omega))$ weak implies that of $\overline{q}^n(0,t)$ to $q(0,t)$ in $L^2(0,T)$ weak. The proof of Theorem 3.2 is now complete.

For uniqueness, we have only the following result.

THEOREM 4.8. *Let us assume that the data are smooth and satisfy the necessary compatibility conditions. Then,* (2.2)–(2.7) *has at most one solution* (p,u) *such that* u *belongs to* $C^\alpha(\overline{Q})$ *with* $\alpha > 0$.

Proof. Note first that if (p,u) is a solution of problem (2.2)–(2.7) with u in $C^\alpha(\overline{Q})$, $\alpha > 0$, then it is well known that p belongs to $C^{2+\alpha,1+\alpha/2}(\overline{Q})$ whenever the data are smooth and satisfy the necessary compatibility conditions. Now let (p,u) and $(\tilde{p},\tilde{u})$ be two solutions, with u and $\tilde{u}$ in $C^\alpha(\overline{Q})$, $\alpha > 0$. Setting $q = -\partial_x p$, $\tilde{q} = -\partial_x \tilde{p}$ and using equations of p and $\tilde{p}$, one can find a constant C such that

$$||\partial_t p - \partial_t \tilde{p}||^2_{L^2(Q)} + ||q - \tilde{q}||^2_{L^2(Q)} \le C\, ||u - \tilde{u}||^2_{L^2(Q)}.$$

Using equations satisfied by u and $\tilde{u}$ and the previous estimate, we show also that there is another constant C' such that

$$\int_0^1 |u(x,t) - \tilde{u}(x,t)|^2\, dx \le C' \int_0^t \int_0^1 |u(x,s) - \tilde{u}(x,s)|^2\, dx\, ds$$

for any t, $0 < t < T$. Hence $u = \tilde{u}$ by virtue of Gronwall's lemma, which proves the theorem.

5. A case of existence and uniqueness of a smooth solution. In this section we are concerned with system (2.2)–(2.3), which is provided with the boundary

and initial conditions

$$p(0,t) = p_1, \quad p(1,t) = p_2, \quad t \in]0,T[, \tag{5.1}$$

$$p(x,0) = p_0(x), \quad x \in \Omega, \tag{5.2}$$

$$u(0,t) = u_1(t), \quad t \in]0,T[\tag{5.3}$$

$$u(x,0) = u_0(x), \quad x \in \Omega, \tag{5.4}$$

where p_1 and p_2 are constants. Comparing this problem to (2.2)–(2.7), we note that we have only replaced (2.4) by (5.1). Assume conditions (2.10)–(2.11) are fulfilled and in addition,

$$p_0 \in C^{2,\overline{\alpha}}(\overline{\Omega}), \quad p_0(0) = p_1, \quad p_1(0) = p_2, \quad p_0''(0) = p_0''(1) = 0, \quad -p_0'' \geq 0, \tag{5.5}$$

$$\text{there exists} \quad m > 0 \quad \text{such that} \quad m \leq -p_0', \tag{5.6}$$

$$\begin{aligned} &u_0 \in C^{1,\overline{\alpha}}(\overline{\Omega}), \quad u_1 \in C^{1,\overline{\alpha}/2}([0,T]), \quad \text{with} \\ &u_0(0) = u_1(0), \quad u_1'(0) - p_0'(0)\, u_0'(0) = 0, \end{aligned} \tag{5.7}$$

$\overline{\alpha} > 0$ to be chosen later. We then state the next theorem.

THEOREM 5.1.

(i) *There is a constant β, $0 < \beta \leq \overline{\alpha}$, such that problem* (2.2)–(2.3), (5.1)–(5.7) *is uniquely solvable in the space $C^{2+\beta,1+\beta/2}(\overline{Q}) \times C^{\beta}(\overline{Q})$.*

(ii) *If, in addition, $0 < u_0(x) < 1$, $0 < u_1(t) < 1$, then u lies in $C^{1+\beta/2}(\overline{Q})$.*

The proof requires several steps. The essential point is the following result of Alkhutov and Mamedov [1] , see also Krylov [7]. Consider the boundary value problem

$$\begin{cases} \partial_t V - K(x,t)\, \partial_x^2 V = f, & (x,t) \in Q, \\ V(0,t) = V(1,t) = 0, & t \in]0,T[, \\ V(x,0) = 0, & x \in \Omega. \end{cases} \tag{5.8}$$

Previously, f is in $L^p(Q)$ and K is a given function in $L^\infty(Q)$ such that there is a constant $\alpha > 0$, with

$$\frac{1}{\alpha} \leq K(x,t) \leq \alpha \qquad \text{for almost every} \quad (x,t) \in Q.$$

Then (see Alkhutov and Mamedov [1], p. 480) we have the following lemma.

LEMMA 5.2. *Under the aforementioned hypotheses, there is σ_0, $0 < \sigma_0 < 1$, such that if $p \in [2-\sigma_0, 2+\sigma_0]$, then problem* (5.8) *is uniquely solvable in $W_p^{2,1}(Q)$ for each $f \in L^p(Q)$ where $W_p^{2,1}(Q) = \{v \in L^p(Q);\ v_t, v_x, v_{xx} \in L^p(Q)\}$.*

The first step consists in proving the next proposition.

PROPOSITION 5.3. *Problem* (2.2), (2.3), *and* (5.1)–(5.4) *admits a solution (p,u) in the space $L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega)) \times H^1(Q)$.*

Proof. It suffices to prove the estimate (4.22) with a constant $C_4(\eta)$ which is in fact independent of the parameter η. To do so, first note, by means of the maximum principle and the new boundary condition (5.1), that

$$m \leq q(x,t) \leq M \quad \text{for almost every} \quad (x,t) \in Q.$$

Next, using again the maximum principle, we can show that

$$\partial_x q(x,t) \geq 0 \qquad \text{for almost every} \quad (x,t) \in Q.$$

Then, from (4.25), after integration over $Q_t = \Omega\times]0,t[$, $0 < t < T$ it follows that

$$\frac{1}{2}\int_0^1 |\partial_x v(x,t)|^2\,dx + \frac{1}{2}\int_0^t \overline{q}(1,s)|\partial_x v(1,s)|^2\,ds + \frac{1}{2}\int_0^t\int_0^1 (\partial_x \overline{q}\,|\partial_x v(x,s)|^2)\,dx\,ds$$
$$= \frac{1}{2}\int_0^1 |\partial_x v(x,0)|^2\,dx + \frac{1}{2}\int_0^t \overline{q}(0,s)|\partial_x v(0,s)|^2\,ds + \int_0^t\int_0^1 \partial_x(v(v-1)q\,\overline{q})\,\partial_x v\,dx\,ds.$$

Since the three terms on the left side are positive, we have a similar estimate to (4.22) with a constant independent of η whenever

$$\int_0^t \overline{q}(0,s)\,|\partial_x v(0,s)J|^2\,ds \le C,$$

where C is a constant independent of η. This can be performed as in Lemma 4.6 with m playing the role of η. Note that we have assumed, as is permissible, that $\overline{q}$ and $\partial_x\overline{q}$ verify

$$\overline{q}(x,t) \ge m, \quad \partial_x\overline{q}(x,t) \ge 0 \quad \text{for almost every} \quad (x,t) \in Q.$$

We now use Lemma 5.2 to show the existence of a smoother solution.

PROPOSITION 5.4. *Let $r = 2 + \sigma_0$, with σ_0 given in Lemma 5.2. Then system (2.2)-(2.3) subject to conditions (5.1)–(5.7) admits a solution (p,u) such that*

(i) $p \in W_r^{2,1}(Q)$,

(ii) $\partial_x u \in L^\infty(0,T;L^r(\Omega))$,

(iii) $\partial_t u \in L^r(0,T;L^r(\Omega))$.

Proof. We use the result of Alkhutov and Mamedov [1], for u given in $\mathcal{C}$, and the solution p of problem (2.2), (5.1), and (5.2) satisfies the estimate

$$||p||_{W_r^{2,1}(Q)} \le C,$$

where C is a constant which depends only on the boundary and initial conditions on p. The exponent r depends only on the values of $a(u)$ and not on those of u. In fact, we must show (ii) and (iii). To do so, we proceed as in §4 by introducing the function v solution of

$$\partial_t v + q\,\partial_x v = v(v-1)\,q\,\overline{q}, \quad (x,t) \in Q, \tag{5.9}$$

$$\begin{aligned} &v(0,t) = u_1(t), \quad t \in]0,T[, \\ &v(x,0) = u_0(x)\,e^{p_0(x)}\,[\,(1-u_0(x)) + u_0(x)\,e^{p_0(x)}\,]^{-1}, \quad x \in \Omega. \end{aligned} \tag{5.10}$$

We must estimate $\partial_x v$ and $\partial_t v$ in $L^\infty(0,T;L^r(\Omega))$. This may be obtained as in the previous section by differentiating (5.9) with respect to x and multiplying by $|\partial_x v|^{r-2}\,\partial_x v$. This allows us to bound $\partial_x v$ and $\partial_t v$ in $L^\infty(0,T;L^r(\Omega))$ and therefore $\partial_x u$ and $\partial_t u$ in $L^\infty(0,T;L^r(\Omega))$ and $L^r(0,T;L^r(\Omega))$ according to the relation

$$u = \frac{e^{-p}\,v}{(1-v) + v\,e^{-p}}.$$

Proposition 5.4 is proved.

Proof of Theorem 5.1. We put $\beta = 1 - \frac{2}{r}$ and assume $\beta \le \overline{\alpha}$. In view of Sobolev's imbedding theorem, since u belongs to $W^{1,r}(Q)$, u is in $C^\beta(\overline{Q})$. Therefore, we see that

by using the pressure equation, if follows, according to Theorem 12.2 in Ladyzhenskaya, Solonnikov and Ural'tseva [8], that p belongs to $C^{2+\beta,1+\beta/2}(\overline{Q})$. Next, if we assume that $0 < u_0 < 1$, $0 < u_1 < 1$, we can introduce the function h (as in Lemma 4.4 with $\varepsilon = 0$) solution of the problem

$$\begin{cases} \partial_t h + q\,\partial_x h = -q^2, & (x,t) \in Q, \\ h(x,0) = \ln\left(\dfrac{u_0(x)}{1-u_0(x)}\right) + p_0(x), & x \in \Omega, \\ h(0,t) = \ln\left(\dfrac{u_1(t)}{1-u_1(t)}\right), & t \in]0,T[. \end{cases}$$

By the method of characteristics one can easily prove, since q and $\partial_x q$ are in $C^{\beta/2}(\overline{Q})$, that h is in $C^{1+\beta/2}(\overline{Q})$. The second part of Theorem 5.1 follows from the relation

$$u = \frac{e^{h-p}}{1+e^{h-p}}.$$

This completes the proof.

REFERENCES

[1] Y. A. Alkhutov and I. T. Mamedov, *The first boundary value problem for nondivergence second order parabolic equations with discontinuous coefficients*, Amer. Math. Soc., 59 (1988), pp. 471–495.

[2] Y. Amirat, K. Hamdache, and A. Ziani, *Homogénéisation d'un modèle d'écoulements miscibles en milieu poreux*, Asymptotic Anal., 3 (1990), pp. 77-89.

[3] ———, *Homogenization of a model of compressible miscible flow in porous media*, Boll. Un. Mat. Ital., 5-B (1991), pp. 463–487.

[4] G. Chavent and J. Jaffré, *Mathematical models and finite elements for reservoir simulation*, North-Holland, Amsterdam, 1986.

[5] J. Douglas and J. E. Roberts, *Numerical methods for a model for compressible miscible displacement in porous media*, Math. of Comp., 41 (1983), pp. 441–459.

[6] A. V. Kazhikhov and V. V. Shelukin, *Unique global solution with respect to time of initial boudary value problems for one-dimensional equations of a viscous gas*, PMM 41-2 (1977), pp. 282–291.

[7] N. V. Krylov, *On equations of minimax type in the theory of elliptic and parabolic equations in the plane*, Math. Sb., 81 (123) (1970), pp. 3–22; English transl., Math. USSR Sb., 10 (1970).

[8] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'tseva, *Linear and quasilinear equations of parabolic type*, Nauka, Moscow, 1967; English transl., American Mathematical Society, Providence, RI, 1968.

[9] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod–Gauthier–Villars, Paris, 1969.

[10] A. E. Scheidegger, *The Physics of flow through porous media*, Univ. Toronto Press, Toronto, 1974.

[11] D. Serre, *Existence globale de solutions faibles des équations de Navier-Stokes d'un fluide compressible en dimension* 1, Sem. Collège de France, X (1990).

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 675–693, May 1995

009

ON ASYMPTOTIC SELF-SIMILAR BEHAVIOUR FOR A QUASILINEAR HEAT EQUATION: SINGLE POINT BLOW-UP *

VICTOR A. GALAKTIONOV†

Abstract. The author studies the asymptotic behaviour near a finite blow-up time $t = T$ of solutions to the degenerate quasilinear parabolic equation

$$u_t = (u^\sigma u_x)_x + u^\beta \quad \text{in } \mathbb{R} \times (0, T),$$

where $\sigma > 0$ and $\beta > \sigma + 1$ are fixed constants. These values of parameters σ, β correspond to single point blow-up. The initial function is assumed to be bounded, symmetric, nonincreasing in $|x|$, and compactly supported. It is proved that the rescaled function $f(\xi, t) = (T-t)^{1/(\beta-1)} u(\xi(T-t)^m, t), m = [\beta - (\sigma+1)]/2(\beta-1) > 0$, behaves as $t \to T$ like a nontrivial self-similar profile $\theta(\xi) > 0$ in $\mathbb{R}, \theta(\xi) \to 0$ as $\xi \to \infty$.

Key words. quasilinear heat equation, single point blow-up, asymptotic behaviour, nonconstant self-similar solution

AMS subject classifications. 34E10, 35B40, 35K55, 35K65

1. Introduction and main result.

We consider the Cauchy problem for the one-dimensional quasilinear degenerate parabolic equation:

$$u_t = (u^\sigma u_x)_x + u^\beta \quad \text{for } x \in \mathbb{R}, t > 0, \tag{1.1}$$

$$u(x, 0) = u_0(x) \geq 0 \quad \text{in } \mathbb{R}; \qquad u_0 \not\equiv 0, \quad \sup u_0 < \infty, \tag{1.2}$$

where $\sigma > 0$ and $\beta > 1$ are fixed constants. This problem is a mathematical model of combustion in a medium where both heat conduction coefficient $k(u) = u^\sigma$ and heat source $Q(u) = u^\beta$ depend upon the temperature $u = u(x, t) \geq 0$. Local in time existence of a weak solution, uniqueness and comparison results for (1.1) and (1.2) are well known; see [4]–[6], [25], [27]–[30], and the references therein. The solution $u = u(x, t)$ is a nonnegative continuous compactly supported function, which is the classical one at any point where $u > 0$. If $1 < \beta < \sigma + 3$, then for a given arbitrary initial function $u_0 \not\equiv 0$, the solution blows up in a finite time $T > 0$ so that

$$\sup_{x \in \mathbb{R}} u(x, t) \to \infty \quad \text{as } t \to T. \tag{1.3}$$

T is then called a finite blow-up time. If $\beta \geq \sigma + 3$, then (1.3) holds for any initial function u_0 large enough (cf. [15] and [30, p. 208]). We assume that the solution $u(x, t)$ of (1.1) and (1.2) blows up as $t \to T$, where T depends on the initial function and exponents σ, β. We also assume that

$$\begin{gathered} u_0 = u_0(|x|), u_0 \text{ does not increase,} \\ u_0 \text{ is a compactly supported function,} \\ \sup |(u_0^\sigma)_x| < \infty. \end{gathered} \tag{1.4}$$

* Received by the editors February 17, 1993; accepted for publication November 3, 1993.

† Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Miusskaya Square 4, 125047 Moscow, Russia. Present address, Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

Then by uniqueness and by the maximum principle we have that $u = u(|x|,t) \geq 0$ in $\mathbb{R} \times (0,T)$, $u(|x|,t)$ does not increase in $|x|$ for any fixed $t \in (0,T)$, and

$$\operatorname{supp} u(x,t) \equiv \{x \in \mathbb{R} | u(x,t) > 0\} = (-h(t), h(t)), \tag{1.5}$$

where $h(t) > 0$ for $t \in [0,T)$ is a continuous nondecreasing function. See a full list of references in [25] concerning properties of interfaces corresponding to weak solutions of degenerate nonlinear heat equations of the type (1.1).

For $\beta \geq \sigma+1$, heat interfaces of a general compactly supported solution have been proved to be localized [11] (see also [30, p. 235]). Namely, there exists the constant $\xi_* = \xi_*(\sigma,\beta) > 0$ such that under above hypotheses on any $t \in [0,T)$

$$h(t) \leq h(0) + \xi_* T^m, \qquad m = \frac{\beta - (\sigma+1)}{2(\beta-1)} \geq 0. \tag{1.6}$$

For $\beta = \sigma + 1$ $(m = 0)$ we have $\xi_*(\sigma, \sigma+1) = L_s/2 \equiv \pi(\sigma+1)^{1/2}/\sigma$, and (1.6) yields the best possible upper estimate [13] (see also [14]): $h(t) \leq h(0) + L_s/2$ in $(0,T)$. If $1 < \beta < \sigma + 1$ $(m < 0)$, then the solution is not localized, $u(|x|,t) \to \infty$ as $t \to T$ for any $x \in \mathbb{R}$ [15], [10] (see also [30, p. 383]), and there holds

$$h(t) = \xi_0 (T-t)^m (1 + o(1)) \to \infty \quad \text{as } t \to T \tag{1.7}$$

(cf. [30, p. 248]), where $\xi_0 = \xi_0(\sigma,\beta) > 0$. Above estimates of interfaces of blowing up solutions have been proved by the method of intersection comparison of $u(x,t)$ with explicit blowing up self-similar solutions of (1.1), which will be discussed later.

This paper is devoted to the analysis of the asymptotic behaviour of the solution near a finite blow-up time for the case $\beta > \sigma + 1$ which corresponds to *single point blow-up*; see [18] and [30, Chap. IV]. We prove that under above hypotheses on the initial function, the solution has a space-time structure of nonconstant self-similar solution near a finite blow-up time. The asymptotic stability of the unique nontrivial symmetric self-similar solution $u_*(x,t) = (T-t)^{-1/\sigma}\theta(x)$ of (1.1) with $\beta = \sigma + 1$ (regional blow-up) has been proved in [12]; see also [30, p. 243].

1.1. Self-similar solution for $\beta > \sigma + 1$. Equation (1.1) admits a blowing up self-similar solution of the form

$$u_*(x,t) = (T-t)^{-1/(\beta-1)}\theta(\xi), \qquad \xi = |x|/(T-t)^m, \tag{1.8}$$

where the function $\theta \geq 0$ satisfies the following nonlinear ordinary differential equation derived by substituting (1.8) into (1.1):

$$A(\theta) \equiv (\theta^\sigma \theta')' - m\theta'\xi - \frac{1}{\beta-1}\theta + \theta^\beta = 0 \quad \text{for } \xi > 0, \tag{1.9}$$

and the symmetry boundary condition

$$\theta'(0) = 0 \ (\theta(0) > 0). \tag{1.10}$$

Problem (1.9), (1.10) with $\beta > \sigma + 1$ admits a positive strictly decreasing solution [1] (see also [30, p. 190] for N-dimensional analysis of a similar equation) having the following behaviour at infinity:

$$\theta(\xi) = C_* \xi^{-2/(\beta-(\sigma+1))}(1 + o(1)) \to 0 \quad \text{as } \xi \to \infty, \tag{1.11}$$

where the constant $C_* > 0$ depends on exponents σ and β. Notice that any solution $\theta(\xi) \geq 0$ in $\mathbb{R}_+$ satisfies (1.11) with some constant $C_* > 0$. Hence, (1.8) is the classical blowing up solution defined in $\mathbb{R} \times [0, T)$. Any solution of (1.9) that satisfies (1.10) and (1.11) is uniformly bounded:

$$\sup_{\xi \in \mathbb{R}} \theta(\xi) < \left(\frac{\beta + \sigma + 1}{(\beta - 1)(\sigma + 2)} \right)^{1/(\beta-1)}$$

(cf. [1] and [30, p. 187]). It follows from (1.8) and (1.11) that for $x \neq 0$

$$u_*(x,t) \to C_* |x|^{-2/(\beta-(\sigma+1))} \quad \text{as } t \to T. \tag{1.12}$$

Moreover, solution (1.8) satisfies $u_t > 0$ in $\mathbb{R} \times [0, T)$ [30, p. 190], and hence

$$u_*(x,t) < C_* |x|^{-2/(\beta-(\sigma+1))} \quad \text{in } (\mathbb{R} \backslash \{0\}) \times [0, T). \tag{1.13}$$

Therefore, $u_*(x,t)$ blows up at the single point $x = 0$. Uniqueness of self-similar solution (1.8) with monotone for $\xi > 0$ function θ remains an open problem. Notice that in general problem (1.9) and (1.10) has at least $M = -[-1/2m] - 1$ ($[p]$ denotes the entire part of p) different positive solutions $\theta_1, \theta_2, \ldots, \theta_M$, where $\theta_k(\xi)$ has exactly k maxima and minima for $\xi \geq 0$; see [1] and [30, p. 188]. The problem of the structure of stable manifolds corresponding to such nonmonotone solutions θ_k is open now.

1.2. The main result. We prove that the asymptotic behaviour as $t \to T$ of the solution $u(x,t), \beta > \sigma + 1$, is described by a blowing up self-similar solution (1.8) having the same blow-up time T. By using a standard technique, we now state the above problem as follows. Let us introduce the new rescaled function:

$$f(\xi, \tau) = (T - t)^{1/(\beta-1)} u(\xi (T - t)^m, t), \tag{1.14}$$

where $\tau = -\log(T - t)$: $[0, T) \to [\tau_0, \infty), \tau_0 = -\log T$, is the new time. Then $f(\xi, \tau)$ satisfies the quasilinear parabolic equation

$$f_\tau = A(f) \quad \text{in } \mathbb{R} \times (\tau_0, \infty), \tag{1.15}$$

$$f(\xi, \tau_0) = f_0(\xi) \equiv T^{1/(\beta-1)} u_0(\xi T^m) \quad \text{in } \mathbb{R}, \tag{1.16}$$

where A is the stationary operator given in (1.9) and f_0 is a nonnegative symmetric compactly supported initial function. Denote by

$$\begin{aligned} \omega(f_0) = \{ & g = g(|\xi|) \geq 0, \quad g \in \mathbb{C}(\mathbb{R}) \mid \exists \tau_j \to \infty \text{ such that} \\ & f(\cdot, \tau_j) \to g(\cdot) \quad \text{as } \tau_j \to \infty \text{ uniformly on compacts in } \mathbb{R} \} \end{aligned} \tag{1.17}$$

the ω-limit set of the solution to the problem (1.15) and (1.16). Therefore the problem of the asymptotic behaviour of $u(x,t)$ near a finite blow-up time $t = T$ is reduced to the problem of stabilization as $\tau \to \infty$ to stationary solutions of equation (1.15).

We now state the main result of the paper.

THEOREM 1.1. *Let $\beta > \sigma + 1$. Assume that* (1.4) *holds. Then*

$$\omega(f_0) \subseteq \{ \theta = \theta(|\xi|) > 0 \mid \theta(\cdot) \textit{ satisfies } (1.9)\text{–}(1.11) \}. \tag{1.18}$$

If problem (1.9)–(1.11) admits a unique, strictly monotone solution $\theta > 0$ (we expect that this is true in one-dimensional case), then (1.18) implies that

$$f(\xi,\tau) \to \theta(\xi) \quad \text{as} \quad \tau \to \infty$$

uniformly on compacts in ξ. Notice also that since $\omega(f_0) \neq \varnothing$ by definition, (1.18) by itself yields the existence of at least one nonnegative weak solution to the stationary problem (1.9) and (1.10).

The main difficulties in the proof of Theorem 1.1 are as follows. The first one is to prove that the constant profile $\theta \equiv (\beta-1)^{-1/(\beta-1)}$ satisfying (1.9) does not appear in the ω-limit. On the other hand, the second difficulty is the fact that degenerate equation (1.15) and (1.9) is not of divergence form, and any explicit Lyapunov function with good properties cannot be constructed. We also need some additional sharp upper and lower estimates of the solution near $t = T$. See §§2–5.

Finally, we note that in the semilinear equation (1.1) with $\sigma = 0$ the corresponding ω-limit set consists of the unique constant profile $\theta \equiv (\beta-1)^{-1/(\beta-1)}$. This has been proved in [19] (see also [20]) by deriving certain upper and lower bounds of solutions and using the construction of an explicit Lyapunov function for the rescaled equation (1.15) and nonexistence result for equation (1.9), $\sigma = 0$, which was proved in [2]. General results on nonexistence of nontrivial self-similar solutions for the N-dimensional equation $u_t = \Delta u + u^\beta$ with $1 < \beta \le (N+2)/(N-2)_+$ and the corresponding formal Lyapunov analysis have been done in [22]. Necessary bounds of a wide class of solutions were proved in [23]. It is interesting that a similar Lyapunov function yields important results about ω-limits for the heat equation with the critical absorption exponent $u_t = \Delta u - u^{1+2/N}$; see [16] and also [30, pp. 121–125]. In this case the choice of a unique, stable nonconstant asymptotic profile is made by using a so-called energy equation for the norm of the solution in the corresponding weighted L^1-space [16], [30, p. 125]. For the quasilinear equation $u_t = \operatorname{div}(u^\sigma \nabla u) - u^{\sigma+1+2/N}$ with $\sigma > 0$, a similar result has been proved in [21] without using a Lyapunov function for the corresponding rescaled equation.

2. First estimates. We begin with simple lower and upper estimates of the solution.

PROPOSITION 2.1 ([30, Chap. IV]). *Assume that* (1.4) *holds. Then*

$$f(0,\tau) > (\beta-1)^{-1/(\beta-1)} = \theta_0 > 0 \quad \textit{for } \tau > \tau_0, \tag{2.1}$$

and there exists a constant $\mu_* > 0$ *such that*

$$f(\xi,\tau) \le \mu_* \quad \textit{in } \mathbb{R} \times (\tau_0,\infty). \tag{2.2}$$

Estimate (2.1) implies that $\omega(f_0)$ does not contain the function $g \equiv 0$. It follows from (2.2) that $\omega(f_0)$ consists of uniformly bounded functions. Notice that $f(\xi,\tau)$ is a compactly supported function for $\tau > \tau_0$ and (1.6) and (1.14) yield

$$\operatorname{supp} f(\xi,\tau) = \{|\xi| < p(\tau)\}, \tag{2.3}$$

where

$$p(\tau) = O(\exp(\tau m)) \to \infty \quad \text{as } \tau \to \infty. \tag{2.4}$$

By using (2.2) and Bernstein-type estimates in the form [3] (cf. general results in [5] and [6]), we obtain the following estimate.

PROPOSITION 2.2. *Under hypotheses* (1.4) *there exists a constant* $\mu_1 > 0$ *such that*

$$|(f^{\sigma+1})_\xi| \le \mu_1 \quad \text{in } \mathbb{R} \times (\tau_0, \infty). \tag{2.5}$$

This implies that the ω-limit set, $\omega(f_0)$, is well defined by (1.17).

The stationary equation

$$A(\theta) = 0 \quad \text{in } \mathbb{R} \tag{2.6}$$

admits the unique nontrivial constant solution $\theta \equiv \theta_0$ in $\mathbb{R}$. The following result implies that $\{g \equiv \theta_0\} \cap \omega(f_0) = \varnothing$.

LEMMA 2.3. *Assume that* (1.4) *holds. Then there exist constants* $C_0 > 0$ *and* $\gamma_0 > \theta_0$ *such that as* $\tau \to \infty$

$$f(\xi, \tau) \le C_0 |\xi|^{-2/(\beta-(\sigma+1))} \quad \text{for } |\xi| > 0, \tag{2.7}$$

$$f(0, \tau) \ge \gamma_0. \tag{2.8}$$

Proof. It was proved in [18] by using the beautiful idea from [8] that under hypotheses (1.4) and the following additional assumption on the initial function

$$u_0^\sigma u_0' + \lambda_0 x u_0^\beta \le 0 \quad \text{in } \mathbb{R}_+, \tag{2.9}$$

where $\lambda_0 \in (0, \lambda_*]$, $\lambda_* = \sigma/(\sigma+2\beta) < 1$, is an arbitrary constant, there holds $J(x,t) = u^\sigma u_x + \lambda_0 x u^\beta \le 0$ in $\mathbb{R}_+ \times (0,T)$. Consider now arbitrary initial data satisfying (1.4). Fix $t_0 \in (0,T)$ and $\delta > 0$ small enough. Since $\beta > \sigma + 1$, we have that $u^{\sigma-\beta} u_x < 0$ is not integrable in x on $(h(t) - \delta, h(t))$. Therefore, for any $t \in [t_0, T)$ there exists $\xi_\delta(t) \in (h(t)-\delta, h(t))$ such that $J(\xi_\delta(t), t) < 0$. Since by the strong maximum principle [7] $(u^{\sigma+1})_{xx}(0, t_0) < 0$, we then conclude that $J(x, t_0) \le 0$ in $(0, \xi_\delta(t_0))$ provided that $\lambda_0 > 0$ is small. Hence it follows from [18] that $J \le 0$ in $(0, h(t) - \delta) \times (t_0, T)$. Since $\delta > 0$ is small, integrating this inequality over $(0, x)$ yields (2.7) with

$$C_0 = \left[\frac{\beta - (\sigma+1)}{2} \lambda_0 \right]^{-1/(\beta-(\sigma+1))}$$

Since $J(0,t) \equiv 0$ we have that $J_x(0,t) \equiv (u^\sigma u_x)_x + \lambda_0 u^\beta \le 0$. This implies that $u_t \le (1 - \lambda_0) u^\beta$ for $x = 0$, and integrating over (t, T) leads to (2.8) with $\gamma_0 = (1 - \lambda_0)^{-1/(\beta-1)} \theta_0$. □

From (2.7) and (2.8) we conclude that if $g \in \omega(f_0)$, then

$$g(0) > \theta_0 \quad \text{and} \quad g(\xi) \to 0 \quad \text{as } |\xi| \to \infty. \tag{2.10}$$

The proof of Theorem 1.1 is based on the construction of a suitable Lyapunov function corresponding to (1.15).

3. Lyapunov function. Set

$$v(\xi, \tau) = f^{\sigma+1}(\xi, \tau). \tag{3.1}$$

Then $v(\xi,\tau) \geq 0$ solves the equation

$$v_\tau = a(v)v_{\xi\xi} + b(\xi, v, v_\xi) \quad \text{in } \mathbb{R} \times (\tau_0, \infty), \tag{3.2}$$

$$v(\xi, 0) = v_0(\xi) \equiv f_0^{\sigma+1}(\xi) \quad \text{in } \mathbb{R}, \tag{3.3}$$

where the coefficients

$$a(v) = |v|^{\sigma/(\sigma+1)}, \tag{3.4}$$

$$b(\xi, v, w) = -m\xi w - \frac{\sigma+1}{\beta-1}v + (\sigma+1)|v|^{(\beta-1)/(\sigma+1)}v, \tag{3.5}$$

are now defined for arbitrary values $v \in \mathbb{R}$. We shall use a general approach to the construction of the Lyapunov function which is due to [31].

For fixed $\xi_0 \geq 0, v_0, w_0$, denote by $\varphi(\xi_0, \xi, v_0, w_0)$ the solution to the ordinary differential equation

$$a(\varphi)\varphi''_{\xi\xi} + b(\xi, \varphi, \varphi'_\xi) = 0 \tag{3.6}$$

either for $\xi \in [0, \xi_0]$ or for $\xi \in [\xi_0, \infty)$, with boundary conditions

$$\varphi|_{\xi=\xi_0} = v_0, \qquad \varphi'_\xi|_{\xi=\xi_0} = w_0. \tag{3.7}$$

Equation (3.6) is a degenerate one. Therefore, formally we have no automatically good properties (e.g., existence, uniqueness, continuous dependence upon parameters, etc.) of the solution that we need for the construction of a Lyapunov function. According to [31], the existence of a Lyapunov function will depend on the aforementioned properties of solutions φ.

3.1. Good and bad properties of functions φ. We begin with some simple preliminary good properties of solutions to (3.6) and (3.7). We shall denote by $\varphi'(\xi_0, \xi, v_0, w_0)$ the derivative $\varphi'_\xi(\xi_0, \xi, v_0, w_0)$.

PROPOSITION 3.1 (see [1] and [30, Chap. IV]). *For any fixed $\xi_0 \geq 0, v_0, w_0$, there exists a weak solution $\varphi(\xi_0, \xi, v_0, w_0)$ of (3.6) and (3.7), a $\mathbb{C}^1$-function for $\xi \in [0, \infty)$, which is bounded with its derivative $\varphi'(\xi_0, \xi, v_0, w_0)$ on any compact in ξ, and $\varphi \in \mathbb{C}^\infty$ at any point where $\varphi \neq 0$.*

Thus, any local in ξ weak solution of (3.6) can be extended for the half-space $[0, \infty)$. Notice that by well-known Bernstein estimates, if $|\varphi| \leq c_1$ on a compact $K = [\xi_1, \xi_2]$ then $|\varphi'|_{\xi_1}^{\xi_2}| \leq c_2$ on K, where the constant c_2 depends on c_1, K, σ, and β. By dividing (3.6) over $a(\varphi)$ and integrating over (ξ_1, ξ_2), we have that $|\varphi'|_{\xi_1}^{\xi_2}|$ is small provided that $|\varphi|$ is small on K; see also [1] and estimates in [17].

The following result is the straightforward consequence of the structure of degenerate equation (3.6) (cf. a similar result in [17, p. 169] for a different degenerate equation).

PROPOSITION 3.2. *Fix some $\xi_0 \geq 0, v_0, w_0$ and an arbitrary compact $K \subset \bar{\mathbb{R}}_+ = [0, \infty)$, such that $\xi_0 \in K$. Assume that the solution $\varphi(\xi_0, \xi, v_0, w_0)$ of (3.6) and (3.7) satisfies*

$$\varphi^2 + (\varphi')^2 \neq 0 \quad \text{for any } \xi \in K. \tag{3.8}$$

Then the functions $\varphi(\xi_0, \xi, v_0, w_0)$ and $\varphi'(\xi_0, \xi, v_0, w_0)$ are continuous on K with respect to small perturbations of the parameters ξ_0, v_0, w_0.

Since (3.6) is degenerate, the continuous dependence on parameters and the uniqueness of the solution to (3.6), (3.7) on K may not exist if $v_0^2 + w_0^2 = 0$, or, in general, if (3.8) is not valid. But it is important that in any case we have the uniqueness to the left of $\xi = \xi_0 > 0$.

PROPOSITION 3.3. *For any given* $\xi_0 > 0$ *and* $v_0 = w_0 = 0$ *there exists the following unique solution to* (3.6) *and* (3.7) *for* $\xi < \xi_0$:

$$\varphi(\xi_0, \xi, 0, 0) \equiv 0 \quad \text{for } \xi \in [0, \xi_0]. \tag{3.9}$$

This is easily proved by a local analysis of (3.6) in a small left neighbourhood of the point $\xi = \xi_0$; see [1], [9], and [24].

We now state some bad properties of "nonuniqueness to the right" to solutions of the degenerate nonlinear equation (3.6).

PROPOSITION 3.4. *For any fixed* $\xi_0 > 0$ *and* $v_0 = w_0 = 0$, *problem* (3.6)–(3.7) *for* $\xi > \xi_0$ *has the unique solution* $\varphi_*(\xi_0, \xi, 0, 0) > 0$ *in a small right neighbourhood of the point* $\xi = \xi_0$.

The proof is based on using the Banach contraction principle applied to the integral equation being equivalent to (3.6) and (3.7) with $v_0 = w_0 = 0$ in a small right neighbourhood of $\xi = \xi_0$. See a similar analysis in [24] and [9] of equations with a degeneracy of the type as given in (3.6). The uniqueness result in Proposition 3.3 is the direct consequence of the structure of the integral equation previously mentioned. A local analysis of the integral equation yields the following behaviour of the solution φ_* for $\xi > \xi_0$:

$$\varphi_*(\xi_0, \xi, 0, 0) = (m\sigma\xi_0(\xi - \xi_0))^{(\sigma+1)/\sigma}(1 + o(1)) \quad \text{as } \xi \to \xi_0 + 0. \tag{3.10a}$$

In the case $\xi_0 = 0$ from the previous integral equation we have the estimate of a solution which could be nonunique,

$$\varphi_*(0, \xi, 0, 0) \le (\text{const } \xi^2)^{(\sigma+1)/\sigma} \quad \text{as } \xi \to +0. \tag{3.10b}$$

By Proposition 3.3, we have that $\varphi_*(\xi_0, \xi, 0, 0) \equiv 0$ for any $\xi \in [0, \xi_0]$.

Thus, for any given $\xi_0 > 0$ problem (3.6) and (3.7) with $v_0 = w_0 = 0$ has the following one-dimensional family of different solutions:

$$\varphi(\xi_0, \xi, 0, 0) = 0 \quad \text{for } 0 \le \xi \le \xi_*, \tag{3.11}$$

$$\varphi(\xi_0, \xi, 0, 0) = \pm\varphi_*(\xi_*, \xi, 0, 0) \quad \text{for } \xi > \xi_*, \tag{3.12}$$

where $\xi_* \ge \xi_0$ is an arbitrary constant. We now sum up bad properties which we will need later.

PROPOSITION 3.5. *There exists a maximal set* $B_* \subset \mathbb{R}^3_+ = \{\xi_0 > 0, v_0 \in \mathbb{R}, w_0 \in \mathbb{R}\}$ *of parameters* (ξ_0, v_0, w_0) *such that for any point* $S \equiv (\xi_0, v_0, w_0) \in B_*$ *the solution* $\varphi(\xi_0, \xi, v_0, w_0)$ *satisfies*

$$\exists \xi_* \in [0, \xi_0] \text{ such that (cf. (3.8))}$$
$$(\varphi(\xi_0, \xi_*, v_0, w_0))^2 + (\varphi'(\xi_0, \xi_*, v_0, w_0))^2 = 0, \tag{3.13}$$

and hence by Proposition 3.3

$$\varphi(\xi_0, \xi, v_0, w_0) \equiv 0 \quad \text{for } \xi \in [0, \xi_*]. \tag{3.14}$$

Notice that, by Proposition 3.2, if $S \notin B_*$ then (3.8) holds on any compact K, and hence there exist uniqueness and continuous dependence of the solution $\varphi(\xi_0, \xi, v_0, w_0)$ on K.

The set B_* that was given also has some good properties.

PROPOSITION 3.6. *Fix arbitrary $(\xi_0, v_0, w_0) \in B_*$. Then, uniformly on $[0, \xi_0]$, the functions $\varphi(\xi_0, \xi, v_0, w_0)$ and φ' are continuous with respect to a small perturbation of parameters ξ_0, v_0, w_0 and in particular*

$$\varphi(\xi_0, \xi, v, w) \to 0 \quad \text{as } v \to v_0, w \to w_0 \tag{3.15}$$

uniformly on $[0, \xi_*]$, where $\xi_* \in [0, \xi_0]$ is given by (3.13).

Proof. First of all, by Proposition 3.2 there exists a continuous dependence of the solution φ and the derivative φ' on any compact subset $K_\varepsilon = [\xi_* + \varepsilon, \xi_0]$ (we now suppose that $0 < \xi_* < \xi_0$), where $\varepsilon > 0$ is an arbitrarily small constant. Suppose on the contrary that there is no continuous dependence of $\varphi(\xi_0, \xi, v_0, w_0)$ on $[0, \xi_*]$. Without loss of generality we may assume that there exist suitable sequences $\{\xi_{0k}\} \to \xi_0, \{v_{0k}\} \to v_0, \{w_{0k}\} \to w_0$ such that the functional sequence $\{\varphi(\xi_{0k}, \xi, v_{0k}, w_{0k})\}$ does not converge to $\varphi(\xi_0, \xi, v_0, w_0) \equiv 0$ as $k \to \infty$ uniformly on $[0, \xi_*]$. Then by standard compactness argument for ordinary differential equations (see Proposition 3.1 and comments given after it), we conclude that there exists a subsequence $\{k'\}$ such that

$$\varphi(\xi_{0k'}, \xi, v_{0k'}, w_{0k'}) \to \bar{\varphi}(\xi) \quad \text{as } k' \to \infty$$

uniformly on $[0, \xi_0]$, where $\bar{\varphi} \in \mathbb{C}^1$ is a weak solution to (3.6) on $[0, \xi_0]$ and $\bar{\varphi} \not\equiv 0$ on $[0, \xi_*]$ by assumption. Since $\bar{\varphi}(\xi) \equiv \varphi(\xi_0, \xi, v_0, w_0)$ on $[\xi_*, \xi_0]$, we have that $\bar{\varphi}(\xi_*) = \bar{\varphi}'(\xi_*) = 0$. Therefore the assertion $\bar{\varphi} \not\equiv 0$ on $[0, \xi_*]$ contradicts Proposition 3.3. Thus, $\bar{\varphi} \equiv 0$ on $[0, \xi_*]$ and that $\varphi' \to 0$ as $k \to \infty$ follows from the gradient estimate stated after Proposition 3.1. This completes the proof. □

3.2. Formal construction of Lyapunov function. Assume now for a moment that (3.2) is a uniformly parabolic equation with smooth bounded coefficients, and suppose that there is no problem with integrabilities at $\xi = \infty$ of given functions on solutions $v(\xi, \tau)$. Then, according to a general approach [31], there exists a formal Lyapunov function of the form

$$L[v](\tau) = \int_{\mathbb{R}} \Phi(\xi, v(\tau), v_\xi(\tau))\, d\xi, \tag{3.16}$$

which is nonincreasing on evolution trajectories corresponding to different solutions $v(\xi, \tau)$:

$$\frac{d}{d\tau} L[v](\tau) = -\int_{\mathbb{R}} \rho(\xi, v, v_\xi)(v_\tau)^2\, d\xi \le 0 \quad \text{for } \tau > \tau_0. \tag{3.17}$$

The functions $\rho(\xi, v, w) \ge 0$ and $\Phi(\xi, v, w)$ are determined as follows.

By using the structure of (3.2) that has a general form, we deduce that (3.17) holds if functions ρ and Φ, assumed to be smooth enough, satisfy the following system of linear partial differential equations:

$$\rho b = -\Phi_v + \Phi_{\xi w} + w\Phi_{vw}, \qquad \rho a = \Phi_{ww}, \tag{3.18}$$

where the coefficients $a(v)$ and $b(\xi, v, w)$ are given in (3.4) and (3.5). Indeed, by formal calculations

$$\frac{d}{d\tau} \int \Phi = \int (\Phi_v v_\tau + \Phi_w v_{\tau\xi}).$$

Integrating by parts in the last term yields

$$\int \Phi_w v_{\tau\xi} = -\int \frac{\partial}{\partial\xi}(\Phi_w)v_\tau = -\int (\Phi_{\xi w} + \Phi_{vw}v_\xi + \Phi_{ww}v_{\xi\xi})v_\tau.$$

Finally, since $v_{\xi\xi} \equiv (v_\tau - b)/a$, we obtain that

$$\frac{d}{d\tau}\int \Phi = \int \left\{ \left[\Phi_v - \Phi_{\xi w} - \Phi_{vw}v_\xi + \frac{b}{a}\Phi_{ww} \right] v_\tau - \frac{1}{a}\Phi_{ww}(v_\tau)^2 \right\}.$$

Hence, (3.17) is valid provided that (3.18) holds. From (3.18) we obtain the linear first-order equation for the function ρ,

$$b\rho_w - aw\rho_v - a\rho_\xi = \rho(wa'_v + a'_\xi - b'_w), \tag{3.19}$$

which can be easily solved by the characteristic's method.

Denote

$$F(\xi, v_0, w_0) = \left. \frac{wa'_v + a'_\xi - b'_w}{a} \right|_{\substack{v=\varphi(0,\xi,v_0,w_0) \\ w=\varphi'(0,\xi,v_0,w_0)}}. \tag{3.20}$$

Then we have

$$\rho(\xi, v, w) = G(v_0, w_0) \exp\left\{ -\int_0^\xi F(\zeta, v_0, w_0)\, d\zeta \right\}\Bigg|_{\substack{v_0=\varphi(\xi,0,v,w) \\ w_0=\varphi'(\xi,0,v,w)}}, \tag{3.21}$$

where G is an arbitrary smooth function to be determined later, and as above $\varphi'(\xi, 0, v, w) = \varphi'_\eta(\xi, \eta, v, w)$ with $\eta = 0$, and

$$\Phi(\xi, v, w) = a\int_0^w (w - \eta)\rho(\xi, v, \eta)\, d\eta + z(\xi, v), \tag{3.22}$$

$$z(\xi, v) = -\int_0^v b(\xi, \mu, 0)\rho(\xi, \mu, 0)\, d\mu. \tag{3.23}$$

It is easily calculated that for coefficients given in (3.4) and (3.5) there holds

$$F(\xi, v_0, w_0) = \frac{\sigma}{\sigma+1}\frac{\varphi'}{\varphi} + m\xi|\varphi|^{-\sigma/(\sigma+1)} \equiv \frac{\sigma}{\sigma+1}(\log|\varphi|)' + m\xi|\varphi|^{-\sigma/(\sigma+1)}, \tag{3.24}$$

where $\varphi = \varphi(0, \xi, v_0, w_0)$. Hence, setting $G(v_0, w_0) = |v_0|^{-\sigma/(\sigma+1)}$, we deduce that

$$\rho(\xi, v, w) = |v|^{-\sigma/(\sigma+1)} \times \exp\left\{ -m\int_0^\xi \zeta|\varphi(0, \zeta, \varphi(\xi, 0, v, w), \varphi'(\xi, 0, v, w))|^{-\sigma/(\sigma+1)}\, d\zeta \right\}. \tag{3.25}$$

It follows from (3.22) and (3.23) that

$$\begin{aligned} \Phi(\xi, v, w) = {} & |v|^{\sigma/(\sigma+1)} \int_0^w (w - \eta)\rho(\xi, v, \eta)\, d\eta \\ & - \int_0^v \rho(\xi, \mu, 0)\left[-\frac{\sigma+1}{\beta-1}\mu + (\sigma+1)|\mu|^{(\beta-1)/(\sigma+1)}\mu \right] d\mu. \end{aligned} \tag{3.26}$$

Notice that formally the Lyapunov function (3.16) satisfies the identity

$$L[v](S) - L[v](\tau_0) = -\int_{\tau_0}^{S} d\tau \int_{\mathbb{R}} \rho(\xi, v, v_\xi)(v_\tau)^2 \, d\xi \le 0 \tag{3.27}$$

for any fixed $S > \tau_0$.

It follows from (3.25), (3.26), and Propositions 3.2 and 3.6 that the functions $\rho(\xi, v, w)$ and $\Phi(\xi, v, w)$ are bounded and continuous for $v > 0$. We set $\rho(0, v, w) = \Phi(0, v, w) = \rho(\xi, 0, 0) = \Phi(\xi, 0, 0) \equiv 0$.

PROPOSITION 3.7. *For any* $\xi \ge 0, v > 0, w \le 0$

$$\rho(\xi, v, w) \le v^{-\sigma/(\sigma+1)}, \tag{3.28}$$

(3.29)

$$\Phi(\xi, v, w) \le \frac{w^2}{2} + (\sigma+1)^2 \left[\frac{1}{(\beta-1)(\sigma+2)} v^{(\sigma+2)/(\sigma+1)} + \frac{1}{\beta+\sigma+1} v^{(\beta+\sigma+1)/(\sigma+1)} \right].$$

In general, these properties are not enough to prove (3.27) for a weak solution of the degenerate equation. We begin with the proof of a weakened form of (3.27) by using some regularizing approach.

4. Regularized problem. It is well known (see [3], [5], [25], [27]–[30], and the references therein) that a weak solution of a nonlinear parabolic equation with a degenerate diffusion operator of nonstationary filtration type can be constructed as the limit of a sequence of classical solutions to a regularized problem. For the Cauchy problem (1.15)–(1.16), this regularized problem can be stated as follows. Fix $\varepsilon > 0$ small enough, and consider the initial-boundary value problem in $B_\varepsilon \times \mathbb{R}_+$, $B_\varepsilon = \{|\xi| < 1/\varepsilon\}$, for quasilinear uniformly parabolic equation (3.2) with the function $a(v)$ replaced by

$$a_\varepsilon(v) = (\varepsilon^2 + v^2)^{\sigma/(2(\sigma+1))}, \tag{4.1}$$

and the boundary and initial data

$$v(\xi, 0) = v_0(\xi) \quad \text{in } B_\varepsilon, \tag{4.2}$$

$$v = 0 \quad \text{for } \xi = \pm 1/\varepsilon, \tau > \tau_0. \tag{4.3}$$

For any small $\varepsilon > 0$ problem (3.2), (4.1)–(4.3) has a unique local in time classical solution $v_\varepsilon = v_\varepsilon(|\xi|, \tau)$, which is strictly monotone in $|\xi|$; see e.g., [7] and [25]. By well-known results (see [5] and [25]) we may conclude that

$$v_\varepsilon(\xi, \tau) \to v(\xi, \tau) \quad \text{as } \varepsilon \to 0 \tag{4.4}$$

uniformly on any compact subset of $\mathbb{R} \times (\tau_0, \infty)$. By general regularity results [5], we also have that

$$(v_\varepsilon)_\xi \to v_\xi \quad \text{as } \varepsilon \to 0 \tag{4.5}$$

in $L^2_{\text{loc}}(\mathbb{R} \times (\tau_0, \infty))$ and uniformly on compact subsets, and, in particular,

$$(v_\varepsilon^\alpha)_\tau \to (v^\alpha)_\tau, \quad \alpha = (\sigma+2)/2(\sigma+1) \quad \text{as } \varepsilon \to 0 \tag{4.6}$$

in $L^2_{\mathrm{loc}}(\mathbb{R}\times(\tau_0,\infty))$. Notice that by this construction at any point of nondegeneracy (ξ,τ), where $v(\xi,\tau)>0$ is the classical solution, we have that as $\varepsilon\to 0$ the function v_ε converges to v with first derivatives. By Bernstein estimates we also have that on any given compact subset of $\mathbb{R}\times(\tau_0,\infty)$ as $\varepsilon\to 0$

$$|v_\varepsilon(\xi,\tau)|\le C_1,\qquad |(v_\varepsilon(\xi,\tau))_\xi|\le C_2.$$

(We shall denote by $C_i>0$ different constants which are independent of $\varepsilon>0$.)

4.1. Lyapunov function for the regularized problem. Since the regularized equation is uniformly parabolic, by the method given in §3 we can construct the classical Lyapunov function. It can be easily calculated by (3.21)–(3.23) with $a(v)$ replaced by the function (4.1). Then instead of (3.25) and (3.26), we obtain the functions

(4.7)
$$\rho_\varepsilon(\xi,v,w)=(\varepsilon^2+v^2)^{-\sigma/(2(\sigma+1))}\times\exp\left\{-m\int_0^\xi\zeta[\varepsilon^2+\varphi_\varepsilon^2(0,\zeta,\varphi_\varepsilon(\xi,0,v,w),\varphi'_\varepsilon(\xi,0,v,w))]^{-\sigma/(2(\sigma+1))}d\zeta\right\},$$

(4.8)
$$\Phi_\varepsilon(\xi,v,w)=(\varepsilon^2+v^2)^{\sigma/(2(\sigma+1))}\int_0^w(w-\eta)\rho_\varepsilon(\xi,v,\eta)\,d\eta - \int_0^v\rho_\varepsilon(\xi,\mu,0)\left[-\frac{\sigma+1}{\beta-1}\mu+(\sigma+1)|\mu|^{(\beta-1)/(\sigma+1)}\mu\right]d\mu.$$

As above, the function $\varphi_\varepsilon(\xi_0,\xi,v_0,w_0)$ is the unique classical solution of the nondegenerate ordinary differential equation

$$a_\varepsilon(\varphi)\varphi_{\xi\xi}+b(\xi,\varphi,\varphi_\varepsilon)=0, \tag{4.9}$$

$$\varphi|_{\xi=\xi_0}=v_0,\qquad \varphi_\xi|_{\xi=\xi_0}=w_0. \tag{4.10}$$

The following upper estimates of $\rho_\varepsilon,\Phi_\varepsilon$ and the lower one of Φ_ε are direct consequences of (4.7) and (4.8).

PROPOSITION 4.1. *For any* $\xi\ge 0, v\ge 0, w\le 0$

$$\rho_\varepsilon(\xi,v,w)\le(\varepsilon^2+v^2)^{-\sigma/(2(\sigma+1))}, \tag{4.11}$$

$$\Phi_\varepsilon(\xi,v,w)\le\frac{w^2}{2}+P_\varepsilon(v), \tag{4.12}$$

where

$$P_\varepsilon(v)=\int_0^v(\varepsilon^2+\mu^2)^{-\sigma/(2(\sigma+1))}\left[\frac{\sigma+1}{\beta-1}\mu+(\sigma+1)\mu^{(\beta+\sigma)/(\sigma+1)}\right]d\mu, \tag{4.13}$$

and

$$\Phi_\varepsilon(\xi,v,w)\ge -R(v), \tag{4.14}$$

where

$$R(v)=\frac{(\sigma+1)^2}{\beta+\sigma+1}v^{(\beta+\sigma+1)/(\sigma+1)}\quad \text{if } v>v_*=\theta_0^{\sigma+1},\qquad R(v)=0\quad \text{if } v\le v_*.$$

We now prove the main estimate following from calculations given above.

LEMMA 4.2. *For any $S > \tau_0$ and $\varepsilon > 0$ small enough, there holds*

$$\int_{\tau_0}^{S} d\tau \int_{B_\varepsilon} \rho_\varepsilon(\xi, v_\varepsilon, (v_\varepsilon)_\xi)(\partial_\tau v_\varepsilon)^2 \, d\xi < C_3, \tag{4.15}$$

where C_3 does not depend on S.

Proof. Denote

$$L_\varepsilon[v](\tau) = \int_{B_\varepsilon} \Phi_\varepsilon(\xi, v, v_\xi) \, d\xi. \tag{4.16}$$

Then by construction,

$$\frac{d}{d\tau} L_\varepsilon[v_\varepsilon](\tau) = -\int_{B_\varepsilon} \rho_\varepsilon(\xi, v_\varepsilon, (v_\varepsilon)_\xi)(\partial_\tau v_\varepsilon)^2 \, d\xi \le 0 \quad \text{for } \tau > \tau_0. \tag{4.17}$$

Integrating over (τ_0, S), $S > \tau_0$, yields

$$L_\varepsilon[v_\varepsilon](S) - L_\varepsilon[v_\varepsilon](\tau_0) = -\int_{\tau_0}^{S} d\tau \int_{B_\varepsilon} \rho_\varepsilon(\xi, v_\varepsilon, (v_\varepsilon)_\xi)(\partial_\tau v_\varepsilon)^2 \, d\xi. \tag{4.18}$$

It follows from (4.18), (1.4), and Proposition 4.1 that

$$\int_{\tau_0}^{S} d\tau \int_{B_\varepsilon} \rho_\varepsilon(\xi, v_\varepsilon, (v_\varepsilon)_\xi)[(v_\varepsilon)_\tau]^2 \, d\xi + L_\varepsilon[v_\varepsilon](S) = L_\varepsilon[v_0] \le C_4. \tag{4.19}$$

Then, using estimate (4.14) in (4.19) yields

$$\int_{\tau_0}^{S} d\tau \int_{B_\varepsilon} \rho_\varepsilon(\xi, v_\varepsilon, (v_\varepsilon)_\xi)(\partial_\tau v_\varepsilon)^2 d\xi \le C_4 + \int_{B_\varepsilon} R(v_\varepsilon(\xi, S)) \, d\xi. \tag{4.20}$$

By (4.4) and (2.2) we have that for large S, provided that $\varepsilon > 0$ is small enough,

$$R(v_\varepsilon(\xi, S)) \le 2\frac{(\sigma+1)^2}{\beta+\sigma+1}\mu_*^{\beta+\sigma+1} \quad \text{for } \xi \in \mathbb{R}. \tag{4.21}$$

Therefore, since by Lemma 2.3 and (4.4) on any compact subset $K \subset \mathbb{R}$,

$$v_\varepsilon(\xi, S) \le 2C_0^{\sigma+1}|\xi|^{-2(\sigma+1)/(\beta-(\sigma+1))} \quad \text{on } K,$$

we conclude that $R(v_\varepsilon(\xi, S)) \equiv 0$ outside some uniformly bounded neighbourhood of the origin $\xi = 0$. Finally, we have

$$\int_{B_\varepsilon} R(v_\varepsilon(\xi, S)) \, d\xi \le C_5. \tag{4.22}$$

Using this estimate in (4.20) implies that

$$\int_{\tau_0}^{S} d\tau \int_{B_\varepsilon} \rho_\varepsilon(\xi, v_\varepsilon, (v_\varepsilon)_\xi)(\partial_\tau v_\varepsilon)^2 \, d\xi \le C_4 + C_5 = C_6, \tag{4.23}$$

where the right-hand side is independent of S. This completes the proof. □

4.2. Passage to the limit $\varepsilon \to 0$. We first need to prove the following estimate.

LEMMA 4.3. *For any fixed $S > \tau_1 = \tau_0 + 1$ and $L > 0$ small enough,*

$$\int_{\tau_1}^{S} d\tau \int_0^L \rho(\xi, v, v_\xi)(v_\tau)^2 \, d\xi \le C_7, \tag{4.24}$$

where the constant C_7 does not depend on S.

We begin with some properties of the solution $\varphi_\varepsilon(\xi_0, \xi, v_0, w_0)$ to the problem (4.9)–(4.10).

PROPOSITION 4.4. *For arbitrary fixed $\xi_0 > 0, v_0 \in \mathbb{R}, w_0 \in \mathbb{R}$ uniformly on $[0, \xi_0]$*

$$\varphi_\varepsilon \to \varphi, \ \varphi'_\varepsilon \to \varphi \quad \text{as } \varepsilon \to 0. \tag{4.25}$$

Proof. The limit (4.25) is the straightforward consequence of the continuous dependence of the solution φ_ε and derivative φ'_ε on the parameter ε provided the limiting solution φ with $\varepsilon = 0$ is such that $\varphi^2 + (\varphi')^2 \neq 0$ on $[0, \xi_0]$ (cf. the similar Proposition 3.2). If φ does not satisfy the aforementioned condition and there exists $\xi_* \in [0, \xi_0]$ such that $\varphi^2 + (\varphi')^2 = 0$ for $\xi = \xi_*$ (hence $\varphi \equiv 0$ in $[0, \xi_*]$), then by using the idea from the proof of Proposition 3.6 (based on the uniqueness result given in Proposition 3.3) we conclude that $\varphi_\varepsilon \to 0 \equiv \varphi, \varphi'_\varepsilon \to 0 \equiv \varphi'$ as $\varepsilon \to 0$ on $[0, \xi_*]$. Here we use a natural gradient bound, see Proposition 3.1 and remarks given after it. Hence, (4.25) holds again, which completes the proof. □

We now introduce the "good" set

$$G_* = \mathbb{R}^3_+ \setminus B_*, \tag{4.26}$$

where the "bad" set B_* is given in Proposition 3.5. It follows from the definition that G_* can be defined also as follows:

$$G_* = \{(\xi_0, v_0, w_0) \in \mathbb{R}^3_+ | \varphi(\xi_0, \xi, v_0, w_0) \text{ satisfies (3.8) on } K = [0, \xi_0]\}. \tag{4.27}$$

Then by continuous dependence of φ' on parameters $(\xi_0, v_0, w_0) \in G_*$ (see Proposition 3.2) we deduce that G_* is open. It follows from (4.27), (3.25), and (3.26) that by standard regularity results

$$\rho \text{ and } \Phi \text{ are smooth enough on } G_* \tag{4.28}$$

and hence, by construction, these functions satisfy (3.18) and (3.19) on G_* in the classical sense. By Propositions 3.7, 4.1, 4.4, and (4.27), we obtain the following.

PROPOSITION 4.5. *As $\varepsilon \to 0$,*

$$(\varepsilon^2 + v^2)^{\sigma/(2(\sigma+1))} \rho_\varepsilon(\xi, v, w) \to v^{\sigma/(\sigma+1)} \rho(\xi, v, w) \quad \text{on } G_*. \tag{4.29}$$

Proof of Lemma 4.3. Fix $L > 0$ small enough and $S > \tau_1$. It follows from (4.4), (4.5), and Propositions 2.1 and 2.2 that for any $(\xi, \tau) \in [0, L] \times [\tau_1, S]$,

$$(\xi, v_\varepsilon, (v_\varepsilon)_\xi) \in [0, L] \times \left[\frac{1}{2}\theta_0^{\sigma+1}, 2\mu_*^{\sigma+1}\right] \times [-2\mu_1, 0] \equiv \mathcal{M}_L.$$

One can see from (3.6) and (3.7) that

$$\mathcal{M}_L \subset G_* \cap \left\{ v \geq \frac{1}{2}\theta_0^{\sigma+1} > 0 \right\}, \tag{4.30}$$

provided that $L > 0$ is small enough. Estimate (4.24) is now a straightforward consequence of (4.15), Proposition 4.5 and (4.4)–(4.6). □

A regularity of the functions ρ and Φ and convergence of the type (4.29) on the bad set B_* are unknown. We have shown only that both functions are continuous on $\mathbb{R}^3_+$; see §3.

5. Proof of Theorem 1.1. Let us rewrite the weighted estimate (4.24) for small $L > 0$:

$$\int_{\tau_1}^{S} d\tau \int_0^L (v^{\sigma/(\sigma+1)}\rho(\xi, v, v_\xi))(\partial_\tau v^\alpha)^2 \, d\xi \leq C_8, \tag{5.1}$$

where $\alpha = (\sigma+2)/2(\sigma+1)$. The function $v^{\sigma/(\sigma+1)}\rho$ in (5.1) given by (3.25) is nonnegative. Unfortunately, it is not strictly positive for $v > 0$. One can see that the following result is valid.

PROPOSITION 5.1. *There holds*

$$\rho(\xi, v, w) = 0 \quad \text{on } B_*, \qquad \rho > 0 \quad \text{on } G_*. \tag{5.2}$$

Proof. If $(\xi, v, w) \in B_*$, then by the definition of B_* there exists $\zeta_0 \in [0, \xi)$ such that $|\varphi(\xi, \zeta, v, w)| \sim (\zeta - \zeta_0)_+^{(\sigma+1)/\sigma}$ as $\zeta \to \zeta_0 > 0$ and $|\varphi(\xi, \zeta, v, w)| \lesssim \zeta^{2(\sigma+1)/\sigma}$ as $\zeta \to 0$; see (3.10)–(3.12). It follows that the integral in the right-hand side of (3.25) diverges, and hence $\rho = 0$. If $(\xi, v, w) \in G_*$, then $\varphi^2 + \varphi'^2 \neq 0$ on $[0, \xi]$ by Proposition 3.5 and (4.27), hence $\rho > 0$. □

By the definition of B_* and Proposition 3.3, we also have the following good property.

PROPOSITION 5.2. *Let $g(\xi) \geq 0$ on a compact $[0, \xi_1]$ be a nonnegative nonincreasing solution of* (1.9) *and* (1.10) *with $g(0) > 0$. Then setting $\bar{g} = g^{\sigma+1}$ we have*

$$B_* \cap \{(\xi, \bar{g}(\xi), \bar{g}'(\xi)) \mid \xi \in (0, \xi_1]\} = \varnothing. \tag{5.3}$$

Since the good set G_* is open, we have the following.

LEMMA 5.3. *Let $g(\xi)$ be as in Proposition* 5.2. *For any point $S_0 = (\xi_0, \bar{g}(\xi_0), \bar{g}'(\xi_0)), \xi_0 \in (0, \xi_1]$, there exists a neighbourhood $Q(S_0) \subset \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_-$ such that*

$$B_* \cap Q(S_0) = \varnothing. \tag{5.4}$$

We now begin with some local version of Theorem 1.1.

LEMMA 5.4. *There exists $\bar{\xi}_1 > 0$ such that*

$$\begin{aligned} \omega(f_0) \subseteq \mathcal{W}(\bar{\xi}_1) \equiv \{ & g \geq 0 \mid g \text{ satisfies } (1.9) \text{ and } (1.10), \\ & g^{\sigma+1} \leq N_* = \mu_*^{\sigma+1} \text{ in } [0, \bar{\xi}_1], \ g^{\sigma+1}(0) \geq n_0 = \theta_0^{\sigma+1}\}. \end{aligned} \tag{5.5}$$

Proof. By using Propositions 2.1 and 2.2, we conclude that for any $\xi_1 > 0$ small enough, the estimates

$$n_0/2 \leq v(\xi, \tau) \leq N_*, \qquad -\mu_1 \leq v_\xi(\xi, \tau) \leq 0 \tag{5.6}$$

hold for $\xi \in [0,\xi_1], \tau > \tau_0$. We derive a lower estimate of $\rho(\xi, v, w)$ in the domain

$$\{0 \le \xi \le \xi_1, n_0/2 \le v \le N_*, -\mu_1 \le w \le 0\} \subset \bar{\mathbb{R}}_+ \times \bar{\mathbb{R}}_+ \times \bar{\mathbb{R}}_-, \tag{5.7}$$

given by estimates (5.6). It follows from the analysis of problem (3.6)–(3.7) considered for $\xi \in [0,\xi_1]$, ξ_1 is small enough, with arbitrary $\xi_0 \in [0,\xi_1]$ and boundary values $v = v_0, w = w_0$ satisfying (cf. (5.6)) $n_0/2 \le v \le N_*, -\mu_1 \le w \le 0$, that for any $\xi \in (0,\xi_1], \zeta \in (0,\xi)$, there holds

$$\varphi(\xi, \zeta, v, w) \ge n_0/4. \tag{5.8}$$

Then we deduce from (3.25) that in the domain given by (5.7),

$$\rho(\xi, v, w) \ge v^{-\sigma/(\sigma+1)} \exp\left\{-\frac{m}{2}\left(\frac{n_0}{4}\right)^{-\sigma/(\sigma+1)} \xi_1^2\right\} \equiv C_9 v^{-\sigma/(\sigma+1)}. \tag{5.9}$$

(Notice that a similar result follows from (4.30), (5.2) since the bounded set $\mathcal{M}_L$ with $L = \xi_1$ is closed and ρ is continuous on $\mathcal{M}_L$.) Using (5.1) with $L = \xi_1$ implies that

$$\int_{\tau_1}^{S} d\tau \int_0^{\xi_1} (\partial_\tau v^\alpha)^2 \, d\xi \le C_{10}, \tag{5.10}$$

where by Lemma 4.3 and (5.9) the right-hand side does not depend on S.

Take now an arbitrary $g \in \omega(f_0)$ so that there exists $\tau_j \to \infty$ such that $v(\xi, \tau_j) \to g(\xi)$ as $j \to \infty$ uniformly on $[0, 2\xi_1]$. Let us rewrite (3.2) with coefficients (3.4) and (3.5) in the form

$$\partial_\tau v^\gamma = \gamma v_{\xi\xi} - m\xi(v^\gamma)_\xi - \frac{1}{\beta - 1} v^\gamma + (v^\gamma)^\beta, \tag{5.11}$$

in $\mathbb{R} \times (\tau_0, \infty)$, where $\gamma = (\sigma + 1)^{-1}$. By Propositions 2.1 and 2.2 and by general regularity results [5], [6] we conclude that

$$v(\xi, \tau_j + s) \to h(\xi, s) \quad \text{as } j \to \infty \tag{5.12}$$

in $L^\infty_{\mathrm{loc}}(\mathbb{R}_+ : C(\mathbb{R}))$, where $h(\xi, s)$ is the unique nonnegative weak solution of the Cauchy problem for the same equation

$$\partial_s h^\gamma = \gamma h_{\xi\xi} - m\xi(h^\gamma)_\xi - \frac{1}{\beta - 1} h^\gamma + (h^\gamma)^\beta, \tag{5.13}$$

which satisfies $h(\xi, 0) = g(\xi)$ in $\mathbb{R}$.

By using a standard technique (cf. [26] and [4]), we have from (5.10) that $h \equiv g$ is a stationary solution to (5.13) on $[0, \bar{\xi}_1]$ with arbitrary $\bar{\xi}_1 < \xi_1$. Indeed, using (5.10) yields for $s \in [0, 1]$

$$\begin{aligned}
&\|v^\alpha(\cdot, \tau_j + s) - v^\alpha(\cdot, \tau_j)\|^2_{L^2((0,\bar{\xi}_1))} \\
&\quad \equiv \int_0^{\bar{\xi}_1} |v^\alpha(\xi, \tau_j + s) - v^\alpha(\xi, \tau_j)|^2 \, d\xi \\
&\quad \le \int_0^{\bar{\xi}_1} d\xi \int_{\tau_j}^{\tau_j + s} |\partial_\tau v^\alpha(\xi, \tau)|^2 \, d\tau \\
&\quad \le \int_0^{\bar{\xi}_1} d\xi \int_{\tau_j}^{\infty} |\partial_\tau v^\alpha(\xi, \tau)|^2 \, d\tau \to 0
\end{aligned}$$

as $j \to \infty$. Hence,

$$
\begin{aligned}
&\|v^\alpha(\cdot,\tau_j+s)-v^\alpha(\cdot,\tau_j)\|^2_{L^2((0,\bar\xi_1)\times(0,1))}\\
&\equiv \int_0^1 \|v^\alpha(\cdot,\tau_j+s)-v^\alpha(\cdot,\tau_j)\|^2_{L^2((0,\bar\xi_1))}\,ds\\
&\le \int_0^{\bar\xi_1} d\xi \int_{\tau_j}^\infty |\partial_\tau v^\alpha(\xi,\tau)|^2\,d\tau \to 0 \quad \text{as } j\to\infty.
\end{aligned}
\tag{5.14}
$$

Therefore, we have that $v^\alpha(\xi,\tau_j+s)\to g^\alpha(\xi)$ as $j\to\infty$ in $L^2((0,\bar\xi_1)\times(0,1))$. Hence, the function $h(\xi,s)$ in (5.12) does not depend on s (see the detailed analysis in [26]), hence the result. □

Proof of Theorem 1.1. We now make a continuation in (5.5) with respect to $\bar\xi_1$. Assume for a contradiction that

$$
X_* = \sup\{\bar\xi_1 > 0 \mid (5.5) \text{ is valid}\} < \infty. \tag{5.15}
$$

Then by continuity we have (cf. (5.5))

$$
\omega(f_0) \subseteq \mathcal{W}(X_*). \tag{5.16}
$$

Then it can be easily seen that any $g\in\omega(f_0)$ is a nonnegative monotone decreasing function on $[0,X_*]$. We now prove that the function

$$
\inf_{\xi\in[0,X_*]} \rho(\xi,v(\xi,\tau),v_\xi(\xi,\tau)) \equiv \rho_*(\tau;X_*) \tag{5.17}
$$

satisfies

$$
\liminf_{\tau\to\infty} \rho_*(\tau;X_*) = C_{11} > 0. \tag{5.18}
$$

Indeed, suppose for a contradiction that (5.18) is false, and hence there exist sequences $\{\xi_k\}\subset[0,X_*]$ and $\{\tau_k\}\to\infty$ such that

$$
\rho(\xi_k,v(\xi_k,\tau_k),v_\xi(\xi_k,\tau_k))\to 0 \quad \text{as } k\to\infty. \tag{5.19}
$$

We can assume that $v(\cdot,\tau_k)\to\bar g(\cdot)\in\mathcal{W}(X_*)$ as $k\to\infty$ uniformly on $[0,X_*]$. Let $\{\xi_k\}\to\xi_0\in[0,X_*]$. Then by compactness we deduce that $v(\xi_k,\tau_k)\to\bar g(\xi_0)$, $v_\xi(\xi_k,\tau_k)\to\bar g'(\xi_0)$ and hence by continuity of the functions ρ and $\bar g(\xi)$ we have that

$$
\rho(\xi_k,v(\xi_k,\tau_k),v_\xi(\xi_k,\tau_k))\to\rho(\xi_0,\bar g(\xi_0),\bar g'(\xi_0))=0. \tag{5.20}
$$

Hence, by Proposition 5.1, $(\xi_0,\bar g(\xi_0),\bar g'(\xi_0))\in B_*$ contradicting (5.3). Thus, (5.18) is valid.

We now prove that there exists a constant $C_{12}>0$ such that

$$
v(\xi,\tau)\ge C_{12} \quad \text{on } [0,X_*] \text{ for large } \tau. \tag{5.21}
$$

If we assume for a contradiction that for some sequence $\{\tau_k\}\to\infty$

$$
\inf_{\xi\in[0,X_*]} v(\xi,\tau_k) \equiv v(X_*,\tau_k)\to 0 \quad \text{as } k\to\infty, \tag{5.22}
$$

then, by compactness, there exists $\bar{g} \in \mathcal{W}(X_*)$ such that $v(\cdot, \tau_k) \to \bar{g}(\cdot)$ uniformly on $[0, X_*]$. By continuity of the heat flux [25] corresponding to the weak solution $v(\xi, \tau)$, it then follows from (5.22) that $v_\xi(X_*, \tau_k) \to 0$. Then by passing to the limit $\tau_k \to \infty$ we arrive at the equalities $\bar{g}(X_*) = \bar{g}'(X_*) = 0$, which contradict the nonexistence of such a solution to the problem (1.9) and (1.10); see Proposition 3.3.

By Proposition 2.2, estimate (5.21) implies that for a small enough fixed $\delta > 0$,

$$v(\xi, \tau) \geq C_{13} \quad \text{in } [0, X_* + \delta] \quad \text{for large } \tau. \tag{5.23}$$

Therefore, since the equation for $v(\xi, \tau)$ is uniformly parabolic on $[0, X_* + \delta]$ for large τ, we conclude that

$$|v_{\xi\xi}(\xi, \tau)| \leq C_{14} \quad \text{in } [0, X_* + \delta] \text{ for large } \tau. \tag{5.24}$$

In particular, this implies that for small fixed $\delta > 0$ any point $(\xi, v(\xi, \tau), v_\xi(\xi, \tau)), \xi \in [X_*, X_* + \delta]$, is contained in a small neighbourhood of the point $(X_*, v(X_*, \tau), v_\xi(X_*, \tau))$ for $\tau > \tau_0$ large enough.

We now prove that there exists small $\delta_0 > 0$ such that (cf. (5.18))

$$\liminf_{\tau \to \infty} \rho_*(\tau; X_* + \delta_0) = C_{15} > 0. \tag{5.25}$$

If it is false, then there exist a decreasing sequence $\{\delta_k > 0\} \to 0$ and $\{\xi_k\} \subset (X_*, X_* + \delta_k)$ (see (5.18)), $\{\tau_k\} \to \infty$, such that (5.19) is valid. As previously seen, we have that $v(\cdot, \tau_k) \to \bar{z}(\cdot) = z^{\sigma+1}, z \in \omega(f_0)$ and by (5.16) there exists $\bar{g} \in \mathcal{W}(X_*)$ such that

$$\bar{z} \equiv \bar{g} \quad \text{on } [0, X_*], \tag{5.26}$$

and by estimates (5.23) and (5.24)

$$\bar{z} \text{ and } \bar{z}' \text{ are continuous at the point } \xi = X_*. \tag{5.27}$$

Since $\{\xi_k\} \to X_* + 0$, by passing to the limit $k \to \infty$ we deduce that by continuity

$$\rho(\xi_k, v(\xi_k, \tau_k), v_\xi(\xi_k, \tau_k)) \to \rho(X_*, \bar{z}(X_*), \bar{z}'(X_*)) = 0,$$

which by (5.26), (5.27) again contradicts Proposition 5.2.

Thus, (5.25) holds for some $\delta_0 > 0$. Hence, $(\xi, v, v_\xi) \in G_*$ for any $\xi \in (0, X_* + \delta_0)$ and τ large enough. It then follows from Proposition 4.5 and (5.1) with $L = X_* + \delta_0$ that (5.10) is valid with $\xi_1 = X_* + \delta$, τ_1 replaced by $\tau_* \gg 1$, and arbitrary $S > \tau_*$. Using the method of the proof of Lemma 5.4 we conclude that $\omega(f_0) \subseteq \mathcal{W}(X_* + \delta_0/2)$, which contradicts the definition of X_* given in (5.15). Hence, $X_* = \infty$, which is equivalent to (1.18), and this completes the proof of Theorem 1.1. □

Acknowledgments. The author thanks Professor R. V. Kohn for helpful discussions at the beginning of this work and is grateful to the Courant Institute of Mathematical Sciences for their hospitality during his visit in May, 1990, when this work was started. This research was completed whilst the author was a visitor at the University of Bristol School of Mathematics. The author is grateful to the Royal Society for their financial support under a Guest Research Fellowship and also for the support of the London Mathematical Society. He also thanks the staff of the School

of Mathematics at the University of Bristol for their hospitality and especially Mrs. R. Rees for typing this paper.

Finally, the author thanks the referee for useful suggestions and remarks.

REFERENCES

[1] M. M. Ad'jutov, Ju. A. Klokov, and A. P. Mikhailov, *Self-similar heat structures with diminishing semiwidth*, Differentsial'nye Uravneniya, 19 (1983), pp. 1107–1114 (in Russian); Differential Equations, 19 (1983), pp. 809–815.

[2] M. M. Ad'jutov and L. A. Lepin, *Nonexistence of blowing up self-similar structures in a medium with source and constant heat conductivity*, Differentsial'nye Uravneniya, 20 (1984), pp. 1279–1281. (In Russian.)

[3] D. G. Aronson, *Regularity properties of flows through porous media*, SIAM J. Appl. Math., 17 (1969), pp. 461–467.

[4] D. G. Aronson, M. G. Crandall, and L. A. Peletier, *Stabilization of solutions of a degenerate nonlinear diffusion problem*, Nonlinear Anal., 6 (1982), pp. 1001–1022.

[5] E. Di Benedetto, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83–119.

[6] ———, *A boundary modulus of continuity of a class of singular parabolic equations*, J. Differential Equations, 63 (1986), pp. 418–447.

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

[8] A. Friedman and B. McLeod, *Blow-up of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425–447.

[9] V. A. Galaktionov, *On some properties of travelling waves in a medium with nonlinear heat conductivity and heat source*, J. Vyshisl. Matem. i Matem. Fiz., 21 (1981), pp. 980–989 (in Russian); USSR Comput. Math. Math. Phys., 21 (1981), pp. 167–176.

[10] ———, *On localization conditions of unbounded solutions to quasilinear parabolic equations*, Dokl. Akad. Nauk. SSR, Ser. Math., 264 (1982), pp. 1035–1040 (in Russian); Soviet Math. Dokl., 25 (1982), pp. 775–780.

[11] ———, *Proof of localization of unbounded solutions of the nonlinear parabolic equation* $u_t = (u^\sigma u_x)_x + u^\beta$, Differentsial'nye Uravneniya, 21 (1985), pp. 15–23 (in Russian); Differential Equations, 21 (1985), pp. 11–18.

[12] ———, *Asymptotic behaviour of unbounded solutions of the nonlinear parabolic equation* $u_t = (u^\sigma u_x)_x + u^{\sigma+1}$, Differentsial'nye Uravneniya, 21 (1985), pp. 1126–1134 (in Russian); Differential Equations, 21 (1985), pp. 751–758.

[13] ———, *Sharp estimate of the support of unbounded solution of a quasilinear degenerate parabolic equation*, Doklady Akad. Nauk. SSR, 309 (1989), pp. 265–268 (in Russian); Soviet Math. Dokl., 40 (1990), pp. 493–496.

[14] ———, *Best possible upper bound for blowup solutions of the quasilinear heat conduction equation with source*, SIAM J. Math. Anal., 22 (1991), pp. 1293–1302.

[15] V. A. Galaktionov, S. P. Kurdyumov, A. P. Mikhailov, and A. A. Samarskii, *On unbounded solutions of the Cauchy problem for parabolic equation* $u_t = \nabla(u^\sigma \nabla u) + u^\beta$, Doklady Akad. Nauk. SSR, 252 (1980), pp. 1362–1364 (in Russian); Soviet Phys. Dokl., 25 (1980), pp. 458–459.

[16] V. A. Galaktionov, S. P. Kurdyumov, and A. A. Samarskii, *On asymptotic "eigenfunctions" of the Cauchy problem for a nonlinear parabolic equation*, Mat. Sb., 126 (1985), pp. 435–472. (In Russian.) English translation: Math. USSR-Sb., 54 (1986), pp. 421–455.

[17] V. A. Galaktionov, S. P. Kurdyumov, S. A. Posashkov, and A. A. Samarskii, *Quasilinear parabolic equation with a complex spectrum of unbounded self-similar solutions*, in *Mathematical Modelling. Processes in Nonlinear Media*, Nauka, Moscow, 1986, pp. 142–182. (In Russian.) English translation: CRC Press Int., to appear.

[18] V. A. Galaktionov and S. A. Posashkov, *On some investigation method for unbounded solutions of quasilinear parabolic equations*, J. Vyshisl. Matem. i Matem. Fiz., 28 (1988), pp. 842–854 (in Russian); USSR Comput. Math. Math. Phys., 28 (1988), pp. 148–156.

[19] ———, *Equation* $u_t = u_{xx} + u^\beta$. *Localization, asymptotic behaviour of unbounded solutions*, preprint, Keldysh Inst. Appl. Math. USSR Acad. Sci., No. 97, Moscow, 1985. (In Russian.)

[20] V. A. GALAKTIONOV AND S. A. POSASHKOV, *Applications of a new comparison theorems for unbounded solutions of nonlinear parabolic equations*, Differentsial'nye Uravnenija, 22 (1986), pp. 1165–1173 (in Russian); Diff. Equat., 22 (1986), pp. 809–815.

[21] V. A. GALAKTIONOV AND J. L. VAZQUEZ, *Asymptotic behaviour of nonlinear parabolic equations with critical exponents. A dynamical systems approach*, J. Funct. Anal., 100 (1991), pp. 435–462.

[22] Y. GIGA AND R. V. KOHN, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297–319.

[23] ———, *Characterising blow up using similarity variables*, Indiana Univ. Math. J., 36 (1987), pp. 1–40.

[24] B. H. GILDING AND L. A. PELETIER, *On a class of similarity solutions of the porous media equation*, J. Math. Anal. Appl., 55 (1976), pp. 351–364.

[25] A. S. KALASHNIKOV, *Some of the qualitative theory of nonlinear degenerate parabolic second-order equations*, Uspekhi Mat. Nauk., 42 (1987), pp. 135–176 (in Russian); Russian Math. Surveys, 42 (1987), pp. 169–222.

[26] M. LANGLAIS AND D. PHILLIPS, *Stabilization of solutions of nonlinear and degenerate evolution equations*, Nonlinear Anal., 9 (1985), pp. 321–333.

[27] O. A. OLEINIK, A. S. KALASHNIKOV, AND CHOU-YU-LIN, *The Cauchy problem and boundary value problems for equations of the type of nonstationary filtration*, Izv. Akad. Nauk. SSR Ser. Math., 22 (1958), pp. 667–704. (In Russian.)

[28] P. E. SACKS, *The initial and boundary value problem for a class of degenerate parabolic equations*, Comm. Partial Differential Equations, 8 (1983), pp. 693–733.

[29] ———, *Global behaviour for a class of nonlinear evolution equations*, SIAM J. Math. Anal., 16 (1985), pp. 233–250.

[30] A. A. SAMARSKII, V. A. GALAKTIONOV, S. P. KURDYUMOV, AND A. P. MIKHAILOV, *Blow-up in Problems for Quasilinear Parabolic Equations*, Nauka, Moscow, 1987. (In Russian.) English translation: Walter de Gruyter, Berlin, 1995.

[31] T. I. ZELENYAK, *On qualitative properties of solutions of quasilinear mixed problems for equations of parabolic type*, Mat. Sb., 104 (1977), pp. 486–510. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 694–714, May 1995

010

TWO-PHASE STEFAN PROBLEM WITH SUPERCOOLING*

IVAN G. GÖTZ† AND BORIS ZALTZMAN‡

Abstract. Both one-dimensional two-phase Stefan problems with the thermodynamic equilibrium condition $\theta(R(t),t)=0$ and with the kinetic rule $\theta(R(t),t)=-\varepsilon\dot{R}(t)$ at the moving boundary $x=R(t)$ are considered. We study the properties of the regular solutions of the problem with equilibrium condition. They are obtained as a limit of solutions of the problem with the kinetic law as $\varepsilon\to 0$. The peculiarity of our problem is the partial supercooling of the liquid phase ($\theta<0$) at the initial state. First, we show that the simply connected supercooled liquid phase disappears in a finite time, and after this the solution becomes the classical one. Second, under appropriate structural assumptions on the initial data, we prove the smoothness of the free boundary $x=R(t)$ everywhere except at a point $t=\bar{t}$. At this point the function R may have a jump $R(\bar{t}+0)-R(\bar{t}-0)>0$ exactly equal to the interval in which $\theta(x,\bar{t}-0)\le -L$. Here L is the dimensionless latent heat.

Key words. kinetic undercooling, supercooled, Stefan problem, blowup

AMS subject classifications. 35K05, 35R35, 80A22

1. Introduction.

The problems considered are based on the usual Stefan model for liquid-solid phase transitions. Let the curve $x=R(t)$ be defined as the interface that separates the liquid phase $\{x>R(t)\}$ from the solid one $\{x<R(t)\}$. We may write the following dimensionless form of the Stefan problem on the space interval (0,1):

$$\theta_t=\theta_{xx} \quad \text{in} \quad Q_T^-\cup Q_T^+, \tag{1.1}$$

where $Q_T^{\pm}=\{(x,t):\pm(x-R(t))>0,\ 0<t<T,\ 0<x<1\}$, subject to the initial and boundary conditions

$$\theta(x,0)=\theta_0(x),\quad 0\le x\le 1, \tag{1.2}$$

$$\theta(i,t)=\theta^i(t),\quad 0\le t\le T,\ i=0,1, \tag{1.3}$$

and the Stefan condition on the interface

$$L\dot{R}(t)=\theta_x(R(t)-0,t)-\theta_x(R(t)+0,t), \tag{1.4}$$

$$R(0)=R_0,\quad 0<R_0<1. \tag{1.5}$$

Here θ is the dimensionless temperature scaled so that the equilibrium phase change temperature is zero, and $L=l/c$, where the constant heat capacity c is assumed to be the same in each phase, and where l is the dimension latent heat.

Assuming the phase equilibrium condition at the phase change interface

$$\theta(R(t),t)=0,\quad 0<t<T, \tag{1.6}$$

*Received by the editors August 19, 1992; accepted for publication (in revised form) October 27, 1993.

†Institute of Applied Mathematics and Statistics, Technical University of Munich, Dachauerstraße 9a, 80335 Munich, Germany. On leave from Lavrent'ev Institute of Hydrodynamics, Siberian Division of the Russian Academy of Sciences, prospect Lavrent'eva 15, 630090, Novosibirsk 90, Russia. The research of this author was supported by a Humboldt Foundation scholarship.

‡Center for Energy and Environmental Physics, Jacob Blaustein Institute for Desert Research, Ben-Gurion University of the Negev, Sede-Boqer Campus 84993, Israel.

we complete the formulation of the problem. If no superheating and supercooling appears, i.e.,

$$\theta(x,t)(x-R(t)) \geq 0 \quad \text{in } Q_T := \{(x,t) : 0 < t < T,\ 0 < x < 1\}, \tag{1.7}$$

then the problem (1.1)–(1.6) has a global classical solution under appropriate smoothness conditions on the initial-boundary data. If the inequality (1.7) does not hold for $t = 0$, or $x = 0, 1$ then a finite time blowup of a solution may appear. (See Sherman [1], Fasano and Primicerio [2], Fasano, Primicerio, Howison, and Ockendon [3] and the references therein.) That means that no global classical solution exists. Therefore we shall mainly use the weak formulation of the problem (1.1)–(1.6).

Problem (A^0). *To find functions* R, θ *which satisfy the boundary conditions* (1.3), *the weak form of the equilibrium condition* (1.6)

$$\theta(x,t) \to 0 \quad \text{as} \quad x \to R(t) \quad \text{for a.e. } t \in (0,T), \tag{1.8}$$

and the following integral identity:

$$-\iint_{Q_T} U(x,t)\psi_t(x,t)dxdt + \iint_{Q_T} \theta_x(x,t)\psi_x(x,t)dxdt = \int_0^1 U_0(x)\psi(x,0)dx, \tag{1.9}$$

for every $\psi \in W_2^{1,1}(Q_T)$, $\psi(i,t) = 0$, $i = 0, 1$, $t \geq 0$, $\psi(x,T) = 0$, $0 < x < 1$; *with*

$$U(x,t) = \theta(x,t) + L \cdot H(x - R(t)), \quad \text{in } Q_T,$$
$$U_0(x) = \theta_0(x) + L \cdot H(x - R_0), \quad 0 < x < 1,$$

where H *is the Heaviside function*

$$H(s) = \begin{cases} 1, & s > 0, \\ 0, & s < 0. \end{cases}$$

Another natural way to complete the problem (1.1)–(1.5) is to introduce the kinetic law of the supercooling and superheating on the melting boundary:

$$\theta^\varepsilon(R^\varepsilon(t),t) = -\varepsilon \dot{R}^\varepsilon(t), \quad 0 < t < T, \tag{1.6$^\varepsilon$}$$

with some positive relaxation parameter ε. (See Visintin [4], Xie [5], Charach, Zaltzman, and Götz [6].)

Problem (A^ε). *To find functions* R^ε, θ^ε *satisfying* (1.1)–(1.5) *and* (1.6)$^\varepsilon$.

Visintin has proved the existence of the weak solution for the problem (A^ε) but with Neumann boundary conditions. He also obtained the solution of the problem (A^0) by letting ε tend to zero. We shall proceed in a similar way.

DEFINITION. *The functions* θ, R *are a regular solution of the problem* (A^0) *if there exists a sequence* $\varepsilon_n \to 0$ *as* $n \to \infty$, *such that the following convergence conditions hold for the solutions* θ^{ε_n}, R^{ε_n} *of the problems* (A^{ε_n}):

$$\begin{aligned} &\theta^{\varepsilon_n} \to \theta \quad \text{weakly in } L_2(0,T;H^1(0,1)), \quad \text{strongly in } L_2(Q_T), \\ &R^{\varepsilon_n} \to R \quad \text{weakly star in } BV(0,T). \end{aligned} \tag{1.10}$$

In §2 we present some preliminary results concerning the existence of solutions and the maximum principle. In the third section we obtain the smoothness (piecewise

with respect to time) of a regular solution of the problem (A^0). Section 4 is devoted to the verification of the inequality (1.7) for a regular solution after some time t_*, although it does not hold at the initial state $t = 0$. That means the disappearance of a supercooled liquid in a finite time t_*. In the fifth section we study the fine structure of a regular solution of the problem (A^0). We show that there exists at most one blowup time $t = \bar{t}$ provided that the initial temperature $\theta_0(x)$ has only one interval in the liquid phase where $\theta_0(x) < -L$. Before the instant $\bar{t}$ and after it a regular solution is smooth. At the point $t = \bar{t}$ the free boundary $R(t)$ may have a jump $R(\bar{t}+0) - R(\bar{t}-0) > 0$ which is exactly equal to the interval where $\theta(x, \bar{t}-0) \le -L$.

2. Preliminary results.

THEOREM 1. *Let the functions* θ_0, θ^i, $i = 0, 1$ *satisfy the smoothness assumptions*

$$\theta^i \in C^1(\mathbf{R}) \cap L_\infty(\mathbf{R}), \quad \theta_0 \in C^1[0, R_0] \cap C^1[R_0, 1] \cap C[0, 1], \tag{2.1}$$

the consistency conditions

$$\theta^i(0) = \theta_0(i), \quad i = 0, 1, \tag{2.2}$$

and suppose that

$$\text{the functions } \theta^0, \ \theta^1 \text{ do not change sign for } t > 0. \tag{2.3}$$

Then the following three statements hold:

1. *There exists a unique solution of the problem* (A^ε) *for every* $\varepsilon > 0$, *for some* $T_\varepsilon > 0$:

$$\begin{aligned} &R^\varepsilon \in C^1(0, T_\varepsilon), \ \theta^\varepsilon \in C(\overline{Q}_{T_\varepsilon}) \cap C^{2,1}(Q^{\pm}_{T_\varepsilon}), \\ &\theta^\varepsilon_x \in C(\overline{Q}^{\pm}_{T_\varepsilon} \setminus \{x = 0, 1\}), \end{aligned} \tag{2.4}$$

$$\text{either } T_\varepsilon = +\infty \text{ or } \min\{1 - R^\varepsilon(T_\varepsilon), R^\varepsilon(T_\varepsilon)\} = 0. \tag{2.5}$$

2. *The following estimates hold independently of* ε:

$$\iint_{Q_T} (\theta^\varepsilon_x)^2 dx dt + \varepsilon \int_0^T |\dot{R}^\varepsilon(t)|^2 dt \le C(T), \tag{2.6}$$

$$\int_0^T |\dot{R}^\varepsilon(t)| dt \le C(T), \tag{2.7}$$

for every bounded $T < T_\varepsilon$.

3. *There exists a regular solution of the problem* (A^0),

$$\theta \in L_2(0, T; H^1(0, 1)) \cap L_\infty(Q_T), \ R \in BV(0, T), \tag{2.8}$$

where $T = \liminf_{\varepsilon_n \to 0} T_{\varepsilon_n}$ *for the corresponding sequence* $\{\varepsilon_n\}$.

The first statement of Theorem 1 is proved in [5]. One can obtain the estimates (2.6) and (2.7) by multiplying the differential equations of the problem (A^ε) by the test functions $\psi_1(x, t) := \theta^\varepsilon(x, t) - f(x, t)$, and $\psi_2(x, t) := \operatorname{sign}\theta^\varepsilon(x, t) - g(x, t)$, respectively,

and then integrating by parts in each subdomain $Q_T^{\pm}$. Here $f(x,t) := \theta^0(t) + (\theta^1(t) - \theta^0(t))x$ and $g(x,t) := sign\theta^0(t) + (sign\theta^1(t) - sign\theta^0(t))x$. Similar estimates were obtained by Visintin in [4], where it also has been proved that the estimates (2.6) and (2.7) imply the convergence conditions (1.10), and consequently the existence of a regular solution of the problem (A^0) with the property (2.8).

LEMMA 1 (maximum principle). *Under the assumptions of Theorem* 1 *the function* θ^ε *may have only positive local minima and only negative local maxima inside of the domain* Q_T. *In particular,*

$$\max\{\sup_{\overline{Q}_T}|\theta^\varepsilon|, \sup_{\overline{Q}_T}|\theta|\} \le \max\{\sup_{(0,1)}|\theta_0|, \sup_{t\in(0,T),i=0,1}|\theta^i(t)|\} =: \theta_M. \tag{2.9}$$

The function θ^ε may have local extrema inside the domain Q_T only on the curve $x = R^\varepsilon(t)$. Therefore the statement of Lemma 1 follows immediately from the Stefan condition (1.4) and from the kinetic law $(1.6)^\varepsilon$ with positive constants L, ε.

LEMMA 2. *Suppose that the strict inequalities*

$$\theta^0(t) < -\gamma, \ \theta^1(t) > \gamma \quad \textit{for } t \ge 0, \quad \textit{for some } \ \gamma > 0, \tag{2.10}$$

hold under the assumptions of Theorem 1. *Then there exist global solutions of the problems* (A^0), (A^ε), *i.e.,* $T = T_\varepsilon = +\infty$. *Moreover,*

$$\eta \le R^\varepsilon(t), R(t) \le 1 - \eta \quad \textit{for } t \ge 0, \quad \textit{for some } \eta > 0. \tag{2.11}$$

Proof. Let us take the linear barrier function

$$u(x) := -\gamma + Ax, \quad x \in (0,1). \tag{2.12}$$

The constant A is chosen so that the function $v := u - \theta^\varepsilon$ is positive for $t = 0$, $x \in (0,1)$ and $t \ge 0$, $x = 0, 1$. Namely,

$$A > \max\{\max_{x\in[0,1]}|\theta_0'(x)|, \sup_{t\ge 0}|\theta^1(t)| + \gamma\}. \tag{2.13}$$

Moreover, we must fulfill the inequality $u(R_0) > 0$, i.e.,

$$A > \frac{\gamma}{R_0}. \tag{2.14}$$

The function v solves the homogeneous heat conduction equation in the subdomains $Q_T^{\pm}$, hence it may have local extrema inside the domain Q_T only for $x = R^\varepsilon(t)$. We suppose that there exists $t_0 > 0$ such that

$$v(x,t) > 0, \ x \in [0,1], \ t \in [0,t_0), \tag{2.15}$$

$$v(R^\varepsilon(t_0), t_0) = 0. \tag{2.16}$$

Then

$$\begin{aligned} u(R^\varepsilon(t_0)) &= \theta^\varepsilon(R^\varepsilon(t_0), t_0) = -\varepsilon\dot{R}^\varepsilon(t_0) \\ &= \frac{\varepsilon}{L}(v_x(R^\varepsilon(t_0) - 0, t_0) - v_x(R^\varepsilon(t_0) + 0, t_0)) < 0. \end{aligned} \tag{2.17}$$

From (2.15) we find the inequality

$$\frac{d}{dt}u(R^\varepsilon(t)) = A\dot{R}^\varepsilon(t) = -\frac{A}{\varepsilon}\theta^\varepsilon(R^\varepsilon(t),t) \geq -\frac{A}{\varepsilon}u(R^\varepsilon(t)) \tag{2.18}$$

for $t \in [0,t_0]$. Because of (2.14), we have $u(R_0) > 0$, and so the differential inequality (2.18) implies that $u(R^\varepsilon(t)) > 0$ must hold for every $t \in [0,t_0]$, which is the same as

$$R^\varepsilon(t) > \eta = \gamma/A, \quad t \geq 0.$$

We complete the proof of Lemma 2 by proving the second part of the inequality (2.11) in a similar way and by using the convergence conditions (1.10). □

3. Smoothness (piecewise with respect to time) of a regular solution of the problem (A^0). Since a regular solution of the problem (A^0) is the limit of the solutions for the problems (A^ε), it is sufficient to prove estimates for the solutions of the problems (A^ε) which are local in time and uniform with respect to $\varepsilon > 0$. The idea of the following statement is similar to that of [7].

LEMMA 3. *In addition to the assumptions of Lemma 2 let us suppose that the following estimates,*

$$\varepsilon_n|\dot{R}^{\varepsilon_n}(\tau)|^2 \leq C, \quad \int_0^1 (\theta_x^{\varepsilon_n}(x,\tau))^2dx \leq C, \tag{3.1}$$

hold for some $\tau \geq 0$ and for every $n \in \mathbf{N}$. Then there exist positive constants ν, C_1 depending on C but independent of ε_n such that

$$\|\dot{R}^{\varepsilon_n}\|_{L_3(\tau,\tau+\nu)} \leq C_1, \tag{3.2}$$

$$\int_\tau^{\tau+\nu}\int_0^1 (\theta_t^{\varepsilon_n})^2dxdt + \max_{t\in[\tau,\tau+\nu]}\int_0^1 (\theta_x^{\varepsilon_n}(x,t))^2dx \leq C_1. \tag{3.3}$$

Proof. Let us define the function $\phi(x,t) = \theta^0(t) + x(\theta^1(t) - \theta^0(t))$. We multiply equation (1.1) by the function $(\theta_t^{\varepsilon_n} - \phi_t)$ and integrate it over the intervals $\Omega^-(t) := (0, R^{\varepsilon_n}(t))$ and $\Omega^+(t) := (R^{\varepsilon_n}(t), 1)$. Summing the results of the integration, we obtain

$$\begin{aligned}\int_0^1 (\theta_t^{\varepsilon_n}(x,t))^2dx + \frac{1}{2}\frac{d}{dt}\left(\int_0^1 (\theta_x^{\varepsilon_n}(x,t))^2dx + \varepsilon_n L(\dot{R}^{\varepsilon_n}(t))^2\right) \\ = -[\theta_x^{\varepsilon_n}]_{R^{\varepsilon_n}(t)+0}^{R^{\varepsilon_n}(t)-0}[(\theta_x^{\varepsilon_n})^2]_{R^{\varepsilon_n}(t)+0}^{R^{\varepsilon_n}(t)-0} + \int_0^1 \theta_t^{\varepsilon_n}\phi_t dx + \phi_{tx}(t)(\theta^1(t) - \theta^0(t)).\end{aligned} \tag{3.4}$$

The third term on the right-hand side is bounded, and the second one, I_2, can be estimated in a standard way:

$$|I_2| \leq \delta\int_0^1 (\theta_t^{\varepsilon_n}(x,t))^2dx + C(\delta) \quad \text{for} \quad \delta > 0. \tag{3.5}$$

Defining $z(t) = \int_0^1 (\theta_x^{\varepsilon_n}(x,t))^2dx$, we can estimate the first term on the right-hand side of the identity (3.4) with the help of the following inequalities:

$$|\theta_x^{\varepsilon_n}(R^{\varepsilon_n}(t) - 0, t)|^3 \leq \int_{x'}^{R^{\varepsilon_n}(t)} |(\theta_x^{\varepsilon_n}(x,t))_x^3|dx + |\theta_x^{\varepsilon_n}(x',t)|^3 \tag{3.6}$$

$$\leq 3\int_{\Omega^-(t)} |\theta^{\varepsilon_n}_{xx}(x,t)||\theta^{\varepsilon_n}_x(x,t)|^2dx + |\theta^{\varepsilon_n}_x(x',t)|^3$$

$$\leq 3\delta\int_{\Omega^-(t)} |\theta^{\varepsilon_n}_{xx}(x,t)|^2dx + \frac{3}{\delta}\int_{\Omega^-(t)} |\theta^{\varepsilon_n}_x(x,t)|^4dx + |\theta^{\varepsilon_n}_x(x',t)|^3$$

$$\leq 3\delta\int_{\Omega^-(t)} |\theta^{\varepsilon_n}_{xx}(x,t)|^2dx$$

$$+\frac{3}{\delta}\left(\int_{\Omega^-(t)} |\theta^{\varepsilon_n}_{xx}(x,t)|^2dx\right)^{1/2}\left(\int_{\Omega^-(t)} |\theta^{\varepsilon_n}_x(x,t)|^2dx\right)^{3/2} + |\theta^{\varepsilon_n}_x(x',t)|^3$$

$$\leq 6\delta\int_{\Omega^-(t)} |\theta^{\varepsilon_n}_{xx}(x,t)|^2dx + \frac{3}{\delta^3}z^3(t) + |\theta^{\varepsilon_n}_x(x',t)|^3,$$

$$|\theta^{\varepsilon_n}_x(R^{\varepsilon_n}(t)+0,t)|^3 \leq 6\delta\int_{\Omega^+(t)} |\theta^{\varepsilon_n}_{xx}(x,t)|^2dx + \frac{3}{\delta^3}z^3(t) + |\theta^{\varepsilon_n}_x(x'',t)|^3, \tag{3.7}$$

where $x' \in (0, R^{\varepsilon_n}(t))$ and $x'' \in (R^{\varepsilon_n}(t), 1)$ are chosen so that

$$|\theta^{\varepsilon_n}_x(x',t)|^2 \leq \int_{\Omega^-(t)} (\theta^{\varepsilon_n}_x(x,t))^2dx/R^{\varepsilon_n}(t) \leq z(t)/\eta, \tag{3.8}$$

$$|\theta^{\varepsilon_n}_x(x'',t)|^2 \leq \int_{\Omega^+(t)} (\theta^{\varepsilon_n}_x(x,t))^2dx/(1-R^{\varepsilon_n}(t)) \leq z(t)/\eta, \tag{3.9}$$

and where the constant $\eta > 0$ is defined as in Lemma 2. Applying the estimates (3.5)–(3.9) to the identity (3.4) for sufficiently small $\delta > 0$ we get

$$\int_0^1 (\theta^{\varepsilon_n}_t(x,t))^2dx + \frac{d}{dt}\left(\int_0^1 (\theta^{\varepsilon_n}_x(x,t))^2dx + \varepsilon_n L(\dot{R}^{\varepsilon_n}(t))^2\right)$$
$$\leq C_3(1 + z^{3/2}(t) + z^3(t)). \tag{3.10}$$

The standard method applied to this differential inequality with the help of the assumptions (3.1) gives us the local estimate

$$\max_{t\in[\tau,\tau+\nu]} z(t) \leq C_4, \tag{3.11}$$

with the constants ν, C_4 independent of ε_n. Furthermore, integrating the inequality (3.10) over the interval $(\tau, \tau+\nu)$, we obtain the estimate (3.3). The estimate (3.2) follows from (3.3), (3.6), and (3.7) coupled with the Stefan condition (1.4). That completes the proof of Lemma 3. □

THEOREM 2. *Under the assumptions of Lemma* 2 *there exists a finite or countable number of intervals* (t_i, t^i), $i \in I$ *such that*

$$|(0,T) \setminus \cup_{i\in I}(t_i,t^i)| = 0 \quad \textit{for all } T > 0, \tag{3.12}$$

$$\sup_{t\in(t_i+\delta,t^i-\delta)} |\dot{R}(t)| + \sup_{x\in(0,1),t\in(t_i+\delta,t^i-\delta)} |\theta_x(x,t)| \leq C^i(\delta), \quad \textit{for all } \delta > 0,\ i \in I, \tag{3.13}$$

where $R(t)$, $\theta(x,t)$ *is a regular solution of the problem* (A^0).

Proof. At first we will show that due to the estimate (2.6) for almost every $\tau \geq 0$ one can choose a subsequence $\varepsilon_n \to 0$ fulfilling the conditions (3.1) of Lemma 3. If a

contrary statement is valid then there exists a set $\mathfrak{T} \in (0,T)$ with nonzero measure such that

$$\int_0^1 |\theta^{\varepsilon_n}(x,\tau)|^2 dx + \varepsilon_n |\dot{R}^{\varepsilon_n}(\tau)|^2 \to +\infty \quad \text{as} \quad n \to \infty$$

for every $\tau \in \mathfrak{T}$ and for every subsequence $\varepsilon_n \to 0$. However, in this case we have

$$\int_0^T \left(\int_0^1 |\theta_x^{\varepsilon_n}(x,\tau)|^2 dx + \varepsilon_n |\dot{R}^{\varepsilon_n}(\tau)|^2 \right) d\tau$$
$$\geq \int_{\tau \in \mathfrak{T}} \int_0^1 |\theta_x^{\varepsilon_n}(x,\tau)|^2 dx + \varepsilon_n |\dot{R}^{\varepsilon_n}(\tau)|^2) d\tau \to +\infty \quad \text{as} \quad n \to \infty,$$

which contradicts (2.6).

By finding the value $\nu(\tau)$ from Lemma 3 for every admissible τ we determine the set $V := \cup_{\tau \geq 0}(\tau, \tau + \nu(\tau))$. One can represent this open set as a unity of either a finite or countable number of nonintersecting intervals $V = \cup_{i \in I}(t_i, t^i)$, which satisfy the property (3.12). Moreover, due to the estimates (3.2) and (3.3), we obtain similar estimates for a regular solution of the problem (A^0):

$$\|\dot{R}\|_{L_3(t_i+\delta, t^i-\delta)} \leq C_i(\delta), \tag{3.14}$$

$$\int_{t_i+\delta}^{t^i-\delta} \int_0^1 \theta_t^2 dx dt + \sup_{t \in (t_i+\delta, t^i-\delta)} \int_0^1 (\theta_x(x,t))^2 dx \leq C_i(\delta) \tag{3.15}$$

for every $\delta > 0$, $i \in I$. From the embedding theorems we have

$$W_3^1(t_i + \delta, t^i - \delta) \subset H^{2/3}[t_i + \delta, t^i - \delta],$$

where $H^{2/3}$ is the space of Hölder-continuous functions with exponent $2/3$. Hence

$$R \in H^{2/3}[t_i + \delta, t^i - \delta] \text{ for all } \delta > 0, \ i \in I. \tag{3.16}$$

The estimate (3.15) implies

$$\|\theta_{xx}(\cdot,t)\|_{L_2((0,R(t))\cup(R(t),1))} = \|\theta_t(\cdot,t)\|_{L_2(0,1)} \leq \text{const}$$

for almost every $t \in (t_i, t^i)$ and therefore from the embedding $W_2^2 \subset H^{1+1/2}$ we obtain

$$\theta(\cdot,t) \in H^{1+1/2}[0,R(t)] \cap H^{1+1/2}[R(t),1] \tag{3.17}$$

for a.e. $t \in (t_i, t^i)$, $i \in I$. We can now apply the results of [8] about solutions of the heat conduction equation with the initial data from $H^{1+\alpha}$ in noncylindrical domains of the class $H^{\frac{1+\alpha}{2}}$. They state

$$\theta \in H^{1+\alpha, \frac{1+\alpha}{2}}(\overline{Q}_{t,\delta}^{\pm}) \quad \text{a.e. } t \in (t_i, t^i), \ \delta > 0, \tag{3.18}$$

where $Q_{t,\delta}^{\pm} := \{(x,\tau) : (x,\tau) \in Q_{t^i-\delta}^{\pm}, \ \tau > t\}$. In our case (see (3.16) and (3.17)), we can choose $\alpha = 1/3$. Then (3.18) and the Stefan condition imply the desired estimate (3.13). That completes the proof of Theorem 2. □

4. Finite time disappearance of the supercooled liquid. Henceforth, in addition to the conditions on the boundary data (2.1), we assume structural restrictions on the initial data, namely, the following:

$$\begin{aligned} &\text{There exists } s_0 \in [R_0, 1) \text{ such that} \\ &\theta_0(x)(x - s_0) > 0,\ x \in [0,1],\ x \neq s_0, R_0. \end{aligned} \tag{4.1}$$

This condition prescribes the phase state of the material in different subintervals of (0,1):

$$\begin{aligned} (0, R_0) &: H(x - R_0) = 0,\ \theta_0(x) < 0 \ - \text{ nonsuperheated solid}, \\ (R_0, s_0) &: H(x - R_0) = 1,\ \theta_0(x) < 0 \ - \text{ supercooled liquid}, \\ (s_0, 1) &: H(x - R_0) = 1,\ \theta_0(x) > 0 \ - \text{ nonsupercooled liquid}. \end{aligned}$$

The properties (2.10) and (2.11) imply that the point $x = 0$ for $t \geq 0$ is always in the nonsuperheated solid phase ($H(0 - R(t)) = 0,\ \theta(0,t) < 0$) and the point $x = 1$ for $t \geq 0$ is always in the nonsupercooled liquid phase ($H(1 - R(t)) = 1,\ \theta(1,t) > 0$). Therefore the supercooled liquid phase lies between the normal solid and liquid phases. The main result of the present section can be formulated briefly as follows: The supercooled liquid phase disappears in a finite time, after that a regular solution of the problem (A^0) coincides with the unique solution of the classical Stefan problem. More precisely, see the next theorem.

THEOREM 3. *Let us assume the fulfillment of* (4.1) *in addition to the conditions of Lemma* 2. *Then there exists* $t_* < t'$ *such that the following inequality,*

$$\theta(x,t)(x - R(t)) \geq 0, \tag{4.2}$$

holds for a.e. $x \in (0,1),\ t \geq t_*$. *Here* $t' := (\frac{7}{2}U_M + 2)/\gamma$ (*see proof of Lemma* 5) *and* $U_M = \theta_M + L$, θ_M *is the maximum absolute value of the temperature* (*see* (2.9)).

Proof of Theorem 3. We consider the functions

$$\begin{aligned} \phi^\varepsilon(x,t) &= \max\{-\theta^\varepsilon(x,t)(x - R^\varepsilon(t)), 0\}, \\ \phi(x,t) &= \max\{-\theta(x,t)(x - R(t)), 0\}. \end{aligned}$$

The statement of Theorem 3 is equivalent to the identity

$$\phi(x,t) \equiv 0,\ x \in (0,1),\ t \geq t_*. \tag{4.3}$$

We use two auxiliary results, which we shall prove later.

LEMMA 4. *Under the conditions of Theorem* 3 *the inequality*

$$\int_0^1 \phi^\varepsilon(x,t_2)dx \leq \int_0^1 \phi^\varepsilon(x,t_1)dx + C\varepsilon \tag{4.4}$$

holds for every $t_1,\ t_2$ *such that* $0 \leq t_1 < t_2 \leq T$.

LEMMA 5. *Under the conditions of Theorem* 3 *there exists* $\varepsilon_0 > 0$ *such that for every* $\varepsilon \in (0, \varepsilon_0)$ *we can choose* $t^\varepsilon \in (0, t')$ *such that*

$$\int_0^1 \phi^\varepsilon(x,t^\varepsilon)dx \leq C\varepsilon, \tag{4.5}$$

where the constant C *is independent of* ε.

Let us continue the proof of Theorem 3. Due to the convergence conditions (1.10) we can state

$$\phi^{\varepsilon_n}(x,t) \to \phi(x,t) \text{ as } \varepsilon_n \to 0 \text{ a.e. in } Q_T. \tag{4.6}$$

Therefore the inequality (4.4) gives us

$$\int_0^1 \phi(x,t_2)dx \le \int_0^1 \phi(x,t_1)dx \tag{4.7}$$

for a.e. $t_1, t_2 \in (0,T)$, $t_2 > t_1$. Moreover, using (4.4) and (4.5), we obtain

$$\begin{aligned} 0 \le \int_0^1 \phi(x,t)dx &= \lim_{\varepsilon_n \to 0} \int_0^1 \phi^{\varepsilon_n}(x,t)dx \\ &\le \lim_{\varepsilon_n \to 0} \left(\int_0^1 \phi^{\varepsilon_n}(x,t^{\varepsilon_n})dx + C\varepsilon_n \right) \le \lim_{\varepsilon_n \to 0} (2C\varepsilon_n) = 0 \end{aligned} \tag{4.8}$$

for a.e. $t \in (t^0, T)$, where $t^0 := \limsup_{\varepsilon_n \to 0} t^{\varepsilon_n}$ and the values t^{ε_n} are given in Lemma 5.

Let us denote

$$t_* := \sup \left\{ t : t \in (0,T),\ \operatorname{ess\,inf}_{0<\tau<t} \int_0^1 \phi(x,\tau)dx > 0 \right\}. \tag{4.9}$$

By (4.8) we have $t_* \le t^0 \le t'$. On the other hand, the convergence (4.6) allows us to choose the sequence $\tau^{\varepsilon_n} \to t_*$ as $n \to \infty$ such that

$$\int_0^1 \phi^{\varepsilon_n}(x,\tau^{\varepsilon_n})dx \to 0 \quad \text{as} \quad n \to \infty.$$

Therefore, letting $n \to \infty$ in the inequality (4.4) with $t_1 = \tau^{\varepsilon_n}$ we finally obtain

$$\int_0^1 \phi(x,t)dx \equiv \phi(y,t) \equiv 0 \quad \text{a.e. } y \in (0,1),\ t \in (t_*, T),$$

which completes the proof of Theorem 3. □

Before starting to prove Lemmas 4 and 5 we study the structure of sets where the functions θ and θ^ε are positive or negative.

LEMMA 6. *Under the assumptions of Theorem* 3 *there exist integrable functions* s^ε, *with* s *satisfying the inequalities*

$$0 \le s(t) \le 1, \quad 0 \le s^\varepsilon(t) \le 1, \quad \text{a.e. } t \in (0,T)$$

such that

$$\theta^\varepsilon(x,t)(x - s^\varepsilon(t)) \ge 0 \quad \text{a.e. in } Q_T, \tag{4.10}$$

$$\theta(x,t)(x - s(t)) \ge 0 \quad \text{a.e. in } Q_T. \tag{4.11}$$

Proof of Lemma 6. Due to the convergence (1.10) it is sufficient to find the functions s^ε for $\varepsilon > 0$. With the help of Sard's theorem (see [10]) one can state that

for almost every $\alpha \in \mathbf{R}$ the level set $\Gamma_{\pm}^{\alpha}$ is a finite sum of smooth curves. Here $\Gamma_{\pm}^{\alpha} = \Gamma^{\alpha} \cap Q_T^{\pm}$, $\Gamma^{\alpha} := \{(x,t) : (x,t) \in Q_T,\ \theta^{\varepsilon}(x,t) = \alpha\}$.

Let us now take an admissible $\alpha > 0$, and an arbitrary point $(\tilde{x}, \tilde{t}) \in \Gamma^{\alpha}$. Now we are going to prove that

$$\theta^{\varepsilon}(x, \tilde{t}) \geq 0 \quad \text{for every } x \in (\tilde{x}, 1). \tag{4.12}$$

In order to do this, we shall construct the continuous function S^{α} such that

$$\theta^{\varepsilon}(S^{\alpha}(t), t) \geq \alpha \ \text{ for } t \in (0, \tilde{t}),\ S^{\alpha}(0) \in (s_0, 1),\ S^{\alpha}(\tilde{t}) = \tilde{x}.$$

Then the inequality $\theta^{\varepsilon} \geq 0$ will hold on the parabolic boundary of the domain $Q^{\alpha} := \{(x,t) :\ t \in (0, \tilde{t}),\ x \in (S^{\alpha}(t), 1)\}$. Using the maximum principle (Lemma 1), we then obtain $\theta^{\varepsilon}(x,t) \geq 0$ for each $(x,t) \in Q^{\alpha}$, which implies the inequality (4.12).

Now we construct the function $S^{\alpha}(t)$. There are two alternatives:

(i) There exists a connected piece $\Gamma \subset (\Gamma_{+}^{\alpha} \cap Q_{\tilde{t}}^{+})$ or $\Gamma \subset (\Gamma_{-}^{\alpha} \cap Q_{\tilde{t}}^{-})$, such that $(\tilde{x}, \tilde{t}) \in \Gamma$.

(ii) There exists a neighborhood V of the point $(\tilde{x}, \tilde{t})$ such that $\theta^{\varepsilon}(x,t) \geq \alpha$ (or $\theta^{\varepsilon}(x,t) \leq \alpha$) for $(x,t) \in V \cap Q_{\tilde{t}}$.

At first we consider the case (i). Using the regularity of the level sets of the solutions for the heat conduction equation (see [8, pp. 178–181]), we find $\hat{t} \in [0, \tilde{t})$ and $S^{\alpha} \in C[\hat{t}, \tilde{t}]$ such that

(1) $\theta^{\varepsilon}(S^{\alpha}(t), t) = \alpha$ for $t \in (\hat{t}, \tilde{t})$, $S^{\alpha}(\tilde{t}) = \tilde{x}$,

(2) $(S^{\alpha}(t), t) \in \Gamma$ for each $t \in (\hat{t}, \tilde{t})$,

(3) $\hat{t} = 0$ or $S^{\alpha}(\hat{t}) = R^{\varepsilon}(\hat{t})$.

If $\hat{t} \neq 0$ then we can afterwards consider the same alternatives (i) or (ii) at the new point $(R^{\varepsilon}(\hat{t}), \hat{t})$.

Now we consider the alternative (ii). Since the solution of the homogeneous heat conduction equation might have minima or maxima only on the boundary of the domain, $\tilde{x} = R^{\varepsilon}(\tilde{t})$ holds. However, using Lemma 1 we obtain

$$\theta^{\varepsilon}(x,t) \geq \alpha \quad \text{on the set } \{x = R^{\varepsilon}(t)\} \cap V \cap Q_{\tilde{t}}.$$

Defining $\hat{t} := \max\{0, \sup\{t :\ \theta^{\varepsilon}(R^{\varepsilon}(t), t) < \alpha,\ t < \tilde{t}\}\}$ and $S^{\alpha}(t) := R^{\varepsilon}(t)$ for $t \in (\hat{t}, \tilde{t})$ we deduce that either $\hat{t} = 0$ or alternative (i) holds at the point $(R^{\varepsilon}(\hat{t}), \hat{t})$, because of choice of $\hat{t}$.

Since the number of the connected pieces of curves Γ_{-}^{α}, Γ_{+}^{α} is finite we complete the construction of the function $S^{\alpha}(t)$ on the interval $(0, \tilde{t})$ by repeating this procedure a finite number of times and consequently prove the validity of the inequality (4.12).

In a similar manner, we can prove that for almost every $\alpha < 0$ and arbitrary $(\tilde{x}, \tilde{t})$ such that $\theta^{\varepsilon}(\tilde{x}, \tilde{t}) = \alpha$ the inequality

$$\theta^{\varepsilon}(x, \tilde{t}) \leq 0 \tag{4.13}$$

holds for each $x \in (0, \tilde{x})$. The validity of the inequalities (4.12) and (4.13) is enough for the existence of the function s^{ε} and, consequently, for the function s. We have completed the proof of Lemma 6. □

Proof of Lemma 4. Let consider the set

$$D_{t_1}^{t_2} := \{(x,t) : t \in (t_1, t_2),\ x \in co(R^{\varepsilon}(t), s^{\varepsilon}(t))\} \ \text{ for } 0 < t_1 < t_2 < T,$$
$$\text{where } co(x_1, x_2) = (\min\{x_1, x_2\}, \max\{x_1, x_2\}) \ \text{ for every } x_1, x_2 \in \mathbf{R}.$$

Due to (4.10) we have $\operatorname{supp}\{\phi\} \subset D_0^T$,

$$\theta^\varepsilon(x,t)(x - R^\varepsilon(t)) \equiv 0 \quad \text{for} \quad (x,t) \in D_0^T \setminus \operatorname{supp}\{\phi\}.$$

Hence

$$\begin{aligned} &\int_0^1 \phi^\varepsilon(x,t_2)dx - \int_0^1 \phi^\varepsilon(x,t_1)dx = \int_0^1 \int_{t_1}^{t_2} \phi_t^\varepsilon(x,t)dxdt \\ &= \iint_{D_{t_1}^{t_2}} \theta_t^\varepsilon(x,t)(R^\varepsilon(t) - x)dxdt + \iint_{D_{t_1}^{t_2}} \theta^\varepsilon(x,t)\dot{R}^\varepsilon(t)dxdt. \end{aligned} \tag{4.14}$$

Moreover, the inequalities (4.10) imply $\theta^\varepsilon(s^\varepsilon(t),t) = 0$, $\theta_x^\varepsilon(s^\varepsilon(t),t) \geq 0$ for $t \in (0,T)$, and so

(4.15)

$$\begin{aligned} &\iint_{D_{t_1}^{t_2}} \theta_{xx}^\varepsilon(x,t)(R^\varepsilon(t) - x)dxdt = \int_{t_1}^{t_2} dt \int_{co(R^\varepsilon(t),s^\varepsilon(t))} \theta_{xx}^\varepsilon(x,t)(R^\varepsilon(t) - x)dx \\ &= \int_{t_1}^{t_2} dt \int_{co(R^\varepsilon(t),s^\varepsilon(t))} \theta_x^\varepsilon(x,t)dx - \int_{t_1}^{t_2} \theta_x^\varepsilon(s^\varepsilon(t),t)|R^\varepsilon(t) - s^\varepsilon(t)|dt \\ &\leq \int_{t_1}^{t_2} \theta^\varepsilon(R^\varepsilon(t),t)\operatorname{sign}(R^\varepsilon(t) - s^\varepsilon(t))dt = \int_{t_1}^{t_2} |\theta^\varepsilon(R^\varepsilon(t),t)|dt. \end{aligned}$$

Summing up (4.14), (4.15), and taking into account that $\theta^\varepsilon(x,t)\dot{R}^\varepsilon(t) \leq 0$ in D_0^T, we obtain

$$\int_0^1 \phi^\varepsilon(x,t_2)dx \leq \int_0^1 \phi^\varepsilon(x,t_1)dx + \int_{t_1}^{t_2} |\theta^\varepsilon(R^\varepsilon(t),t)|dt.$$

Estimating the second integral in the right-hand side of this inequality by using the kinetic condition $(1.6)^\varepsilon$ and the estimate (2.7), we end with

$$\int_{t_1}^{t_2} |\theta^\varepsilon(R^\varepsilon(t),t)dt| = \varepsilon \int_{t_1}^{t_2} |\dot{R}^\varepsilon(t)|dt \leq \varepsilon C,$$

and complete the proof of Lemma 4. □

Proof of Lemma 5. Due to the uniform boundedness of the functions ϕ^ε, it is sufficient to estimate the value $|s^\varepsilon(t) - R^\varepsilon(t)|$. If $s_0 = R_0$ then the inequality (4.5) holds for $t^\varepsilon := 0$ and every $\varepsilon > 0$. Otherwise, if $s_0 > R_0$ then the value

$$t^\varepsilon := \sup\{t : \ s^\varepsilon(t) - R^\varepsilon(t) \geq \varepsilon, \ t \in (0,T)\}$$

is positive for $\varepsilon < s_0 - R_0$. Let us consider the function $r(t) = R^\varepsilon(t) + \varepsilon$ on the interval $(0,t^\varepsilon)$. By the definition of t^ε we have $r(t) \leq s^\varepsilon(t)$ for $t \in (0,t^\varepsilon)$, and

$$\theta^\varepsilon(x,t) \leq 0 \ \text{ in } D := \{(x,t) : \ 0 < x < r(t), \ 0 < t < t^\varepsilon\}. \tag{4.16}$$

We define the test function as follows:

$$\psi(x,t) = \begin{cases} (r(t) - x)/\varepsilon & \text{for } R^\varepsilon(t) < x < r(t), \ t \in (0,t^\varepsilon), \\ x/R^\varepsilon(t) & \text{for } 0 < x < R^\varepsilon(t), \ t \in (0,t^\varepsilon). \end{cases}$$

Testing (1.1) by the function ψ in the domain D we obtain the identity

$$\begin{aligned}(4.17)\quad &\int_0^{r(t)} U^\varepsilon(x,t)\psi(x,t)dx|_{t=0}^{t=t^\varepsilon} - \iint_D U^\varepsilon\psi_t dxdt \\ &= -\iint_D \psi_x\theta_x^\varepsilon dxdt = -\int_0^{t^\varepsilon} \frac{R^\varepsilon(t)+\varepsilon}{R^\varepsilon(t)\varepsilon}\theta^\varepsilon(R^\varepsilon(t),t)dt + \frac{1}{\varepsilon}\int_0^{t^\varepsilon}\theta^\varepsilon(r(t),t)dt \\ &\quad + \int_0^{t^\varepsilon}\frac{\theta^0(t)}{R^\varepsilon(t)}dt,\end{aligned}$$

where $U^\varepsilon(x,t) = \theta^\varepsilon(x,t) + LH(x - R^\varepsilon(t))$ in Q_T. The first integral on the left-hand side of the identity can be estimated by the maximum of the function $|U^\varepsilon(x,t)|$:

$$(4.18)\qquad |I_1| \le 2U_M.$$

Let us estimate the second integral on the left-hand side of the identity (4.17):

(4.19)

$$\begin{aligned}|I_2| \le U_M\iint_D |\psi_t|dxdt = U_M\Bigg(&\int_0^{t^\varepsilon}\int_0^{R^\varepsilon(t)}\left|\frac{x\dot{R}^\varepsilon(t)}{(R^\varepsilon(t))^2}\right|dxdt \\ &+\int_0^{t^\varepsilon}\int_{R^\varepsilon(t)}^{r(t)}\frac{|\dot{R}^\varepsilon(t)|}{\varepsilon}dxdt\Bigg) \le \frac{3}{2}U_M\int_0^{t^\varepsilon}|\dot{R}^\varepsilon(t)|dt \le \frac{3}{2}U_M.\end{aligned}$$

The first integral on the right-hand side of (4.17) can be estimated as follows:

$$(4.20)\qquad |J_1| = \left|\int_0^{t^\varepsilon}\frac{R^\varepsilon(t)+\varepsilon}{R^\varepsilon(t)}\dot{R}^\varepsilon(t)dt\right| \le \left(1+\frac{\varepsilon}{R_0}\right).$$

Because of (4.16) the second integral on the right-hand side of (4.17) is nonpositive, and therefore (4.17)–(4.20) imply

$$\frac{7}{2}U_M + 1 + \frac{\varepsilon}{R_0} \ge -\int_0^{t^\varepsilon}\theta^0(t)/R^\varepsilon(t)dt \ge \int_0^{t^\varepsilon}|\theta^0(t)|dt > \gamma t^\varepsilon.$$

Using the definition of t' and choosing $\varepsilon_0 < R_0$, we obtain $t^\varepsilon < t'$ and $|s^\varepsilon(t^\varepsilon) - R^\varepsilon(t^\varepsilon)| \le \varepsilon$, completing the proof of Lemma 5. □

COROLLARY 1. *If $s_0 = R_0$ then $t_* = 0$.*

This follows immediately from the proof of Lemma 5.

COROLLARY 2. *The function R is nondecreasing on the interval $(0, t_*)$.*

Let us consider the values t^ε, defined in Lemma 5. Obviously, $\liminf_{\varepsilon\to 0} t^\varepsilon \ge t_*$. On the other hand $s^\varepsilon(t) \ge R^\varepsilon(t) + \varepsilon$ for $t \in (0, t^\varepsilon)$, and hence

$$\dot{R}^\varepsilon(t) = -\theta^\varepsilon(R^\varepsilon(t),t)/\varepsilon \ge 0 \quad \text{for } t \in (0, t^\varepsilon).$$

Therefore the function R is monotone nondecreasing on the interval $(0, t_*)$ as the limit of the nondecreasing functions R^ε.

COROLLARY 3.

$$(4.21)\qquad \inf_{0<\tau<t}\left(\liminf_{\tau'\to\tau} s(\tau') - \limsup_{\tau'\to\tau} R(\tau')\right) > 0$$

for every $t \in (0, t_*)$.

Assume the contrary is valid, namely, that there exists $t \in (0, t_*)$ such that

$$\liminf_{\tau \to t} s(\tau) \le \limsup_{\tau \to t} R(\tau). \tag{4.22}$$

The function θ is continuous outside any neighborhood of the free boundary $x = R(t)$ since it is the solution of the heat conduction equation in the domain Q_T^+. Therefore (4.22) implies $\theta(x,t) \ge 0$ for $x > \limsup_{\tau \to t} R(\tau) = R(t+0)$ and so

$$\lim_{\tau \to t+0} \int_0^1 \phi(x,\tau)dx = 0.$$

Then from the inequality (4.7) we immediately conclude that

$$\int_0^1 \phi(x,\tau)dx \equiv 0 \quad \text{for} \quad \tau \in (t, t_*)$$

and obtain the desired contradiction with the definition of t_* (4.9).

5. Fine structure of a regular solution. Now we start to study the behavior of a regular solution up to the time t_*, i.e., in the presence of the supercooled liquid phase. We assume the simplified structure of the initial data as follows:

$$\begin{aligned} &\text{There exist } r_0^+,\ r_0^- \in (R_0, 1), \quad \text{such that} \\ &\theta_0(x) < -L, \quad x \in (r_0^-, r_0^+), \\ &\theta_0(x) > -L, \quad x \in [R_0, r_0^-) \cup (r_0^+, 1]. \end{aligned} \tag{5.1}$$

As we know from [2], the only reason for the blowup of the solution for the one-phase Stefan problem is the presence of the set $\{\theta \le -L\}$. For the sake of simplicity we consider the case when there is only one simply connected component of this critical set.

THEOREM 4. *Let us suppose, in addition to the assumptions of Theorem* 3, *that the structure of the initial data satisfies* (5.1). *Then for a regular solution of the problem* (A^0) *there exists at most one point* $\bar{t} \in [0, t_*]$ *such that*

$$\sup_{t \in (\delta, \bar{t}-\delta) \cup (\bar{t}+\delta, T)} |\dot{R}(t)| \le C(\delta) \quad \textit{for every } \delta > 0. \tag{5.2}$$

Moreover,

$$R(\bar{t}+0) \ge R(\bar{t}-0), \tag{5.3}$$

$$\theta(x, \bar{t}-0) \le -L \textit{ for } x \in (R(\bar{t}-0), R(\bar{t}+0)), \tag{5.4}$$

$$\theta(x, \bar{t}+0) = \theta(x, \bar{t}-0) > -L \textit{ for } x \in (R(\bar{t}+0), 1]. \tag{5.5}$$

Proof. At first we study the structure of the critical set

$$M := \{(x,t) :\ \theta(x,t) < -L,\ R(t) < x < 1,\ 0 < t < t_*\}.$$

We show that, in spite of the nonsmoothness of the boundary $x = R(t)$, the structure of the set $M \cap \{t = \tau\}$ is similar to the structure of the set $M \cap \{t = 0\}$ for every $\tau > 0$.

LEMMA 7. *Under the condition of Theorem* 4 *there exist functions* r^+, r^- *defined on* $(0,t_r)$, *such that*

$$R(t) \le r^-(t) < r^+(t) < 1, \text{ for } t \in (0,t_r), \tag{5.6}$$

$$(x,t) \in M \text{ iff } 0 < t < t_r,\ r^-(t) < x < r^+(t). \tag{5.7}$$

Moreover the set M *is connected.*

Proof of Lemma 7. Using Corollaries 2 and 3, we can choose a function $l \in C^2(0,t_*)$ such that

$$R(t) < l(t) \le s(t),\ t \in (0,t_*),\ l(0) \in (r_0^+, s_0), \tag{5.8}$$

$$-L < \theta(l(t),t) \le 0,\ t \in (0,t_*), \tag{5.9}$$

$$\sup_{0<\tau<t}(l(\tau) - R(\tau)) > \frac{1}{2}\sup_{0<\tau<t}(s(\tau) - R(\tau)),\ t \in (0,t_*). \tag{5.10}$$

Now we study the properties of the function θ as the solution of the heat conduction equation in the domain $\Omega_\tau := \{(x,t) : 0 < t < \tau,\ R(t) < x < l(t)\}$ with the following boundary condition on the nonsmooth boundary $x = R(t)$:

$$\theta(x,t) \to 0 \text{ as } x \to R(t) + 0, \text{ a.e. } t \in (0,t_*). \tag{5.11}$$

Because of Corollary 2 we know only that the function R is nondecreasing on the interval $(0,t_*)$. Let us approximate the function R by nondecreasing functions $R^\delta \in C^2(0,t_*)$ for $\delta \in (0,\delta_0)$ such that

$$R(t) < R^{\delta'}(t) \le R^\delta(t) < 1,\ t \in (0,t_*),\ \delta', \delta \in (0,\delta_0),\ \delta' < \delta; \tag{5.12}$$

$$R^\delta(t) \to R(t),\ \delta \to 0, \text{ a.e. } t \in (0,t_*). \tag{5.13}$$

For every $\delta \in (0,\delta_0)$ we solve the problem

$$\begin{cases} \theta_t^\delta - \theta_{xx}^\delta = 0, \text{ in } \Omega_\tau^\delta := \{(x,t) : t \in (0,\tau), R^\delta(t) < x < l(t)\}, \\ \theta^\delta(R^\delta(t),t) = 0,\ 0 < t < \tau, \\ \theta^\delta(l(t),t) = \theta(l(t),t),\ 0 < t < \tau, \\ \theta^\delta(x,0) = \theta_0^\delta(x),\ R^\delta(o) < x < 1, \end{cases} \tag{5.14}$$

where τ and δ_0 are chosen so that $R^{\delta_0}(t) < l(t)$ for $t \in (0,\tau)$. The function $\theta_0^\delta \in C^2(R^\delta(0), l(0))$ satisfies the conditions

$$\theta_0^\delta(R^\delta(0)) = 0,\ \theta_0^\delta(l(0)) = \theta_0(l(0)), \tag{5.15}$$

$$\theta_0(x) \le \theta_0^{\delta'}(x) \le \theta_0^\delta(x) \le 0 \text{ for } x \in (R^\delta(0), l(0)),\ \delta' < \delta, \tag{5.16}$$

$$\theta_0^\delta(x) \to \theta_0(x) \text{ as } \delta \to 0,\ x \in (R(0), l(0)), \tag{5.17}$$

$$\text{the function } \theta_0^\delta(x) \text{ has the structure (5.1).} \tag{5.18}$$

Conditions (5.15) and (5.16) imply the existence of the solutions θ^δ of problems (5.14) which are continuous in $\overline{\Omega}_\tau^\delta$ and smooth in Ω_τ^δ and which satisfy

$$\theta(x,t) \le \theta^{\delta'}(x,t) \le \theta^\delta(x,t) \le 0 \text{ in } \Omega_\tau^\delta \tag{5.19}$$

for every $\delta', \delta \in (0,\delta_0)$, $\delta' < \delta$.

For arbitrary $t = t_0$, $t_0 \in (0, \tau)$ we consider the level sets

$$G^a := \{(x,t) : \theta^\delta(x,t) = a,\ (x,t) \in \Omega^\delta_{t_0}\},\ a < -L$$

starting with $t = t_0$. Since the function θ^δ is continuous in $\overline{\Omega}^\delta_{t_0}$ and greater than $-L$ on the side boundaries, the simply connected pieces of G^a can be represented for a.e. $a < -L$ by graphs of the functions $g^a \in W_1^1(0, t_0)$ (see, for example, [9, pp. 178–181]), i.e.,

$$G^a = \{(x,t) : x = g^a(t),\ t \in (0, t_0)\}.$$

We consider the domain $D = \{(x,t) : g^a(t) < x < g^{a'}(t),\ t \in (0, t_0)\}$ for every admissible pair $a, a' < -L$ such that $g^a(t) < g^{a'}(t)$ on $(0, t_0)$. Since $\theta^\delta < -L$ on the parabolic boundary of the domain D (because of (5.18)), this inequality holds also inside D, in particular,

$$\theta^\delta(x, t_0) < -L \text{ for } g^a(t_0) < x < g^{a'}(t_0).$$

Since this inequality is valid for a.e. $a, a' < -L$, we conclude that the set $M \cap \{x = t_0\}$ is the interval

$$\{(x, t_0) : \theta^\delta(x, t_0) < -L\} = (r^-_\delta(t_0), r^+_\delta(t_0)).$$

Doing the same thing for every $t_0 \in (0, \tau)$ we define the functions $r^\pm_\delta$ on the whole interval $(0, \tau)$.

Let us take the function $\theta^0(x,t)$—the pointwise limit of the functions $\theta^\delta(x,t)$ as $\delta \to 0$. It exists as a consequence of inequality (5.19). The functions θ^0 and θ solve the same problem in the domain Ω_τ. Moreover, due to (5.19), the values $r^+_\delta(t)$ do not decrease as $\delta \to 0$ and the values $r^-_\delta(t)$ do not increase as $\delta \to 0$, $t \in (0, \tau)$. Defining the functions $r^\pm$ as the pointwise limits of the functions $r^\pm_\delta$, respectively, we obtain

$$\theta^0(x,t) < -L \iff t \in (0, t_*),\ x \in (r^-(t), r^+(t)). \tag{5.20}$$

The set $M^0 := \{(x,t) : \theta^0(x,t) < -L,\ t \in (0, t_*),\ x \in (R(t), 1)\}$ is connected, because $M^0 = \cup_\delta M^\delta$ and the sets $M^\delta := \{(x,t) : \theta^\delta(x,t) < -L,\ t \in (0, t_*),\ x \in (R^\delta(t), 1)\}$ are connected.

Now we need only to prove that the functions θ and θ^0 coincide. The function $v := \theta^0 - \theta$ solves the following homogeneous problem:

$$\begin{cases} v_t - v_{xx} = 0 \ \text{ in } \Omega_\tau, \\ v(x,0) = 0,\ R(0) \le x \le l(0), \\ v(l(t), t) = 0,\ 0 < t < \tau, \\ v(x,t) \to 0 \ \text{ as } x \to R(t) + 0,\ 0 < t < \tau. \end{cases} \tag{5.21}$$

LEMMA 8. *The function $v(x,t) \equiv 0$ is the unique bounded solution of the problem* (5.21).

Proof of Lemma 8. For every positive δ,

$$\delta < \delta_0 := \inf_{0<t<\tau}(l(t) - R(t)),$$

we define the function

$$r^\delta(t) = R(t) + \delta, \quad t \in (0, \tau)$$

and the domain

$$D^\delta = \{(x,t) : t \in (0,\tau),\ x \in (r^\delta(t), l(t))\}.$$

The distance from every point $(x_0,t_0) \in \overline{D}^\delta$ to the left part of the parabolic boundary of the domain

$$\{(x,t) :\ 0 < t \le t_0,\ R(t_0-0) < x < l(t_0)\}$$

is no less than δ. Since the function v satisfies the heat conduction equation in this domain, the local estimates of the solutions for parabolic differential equations imply $v \in W_2^{2,1}(D^\delta)$ for every $\delta \in (0,\delta_0)$. Multiplying the differential equation of the problem (5.21) by the function v, we obtain the following after integrating by parts in the domain D^δ:

$$\frac{1}{2}\int_{R(0)+\delta}^{R(\tau)+\delta} v^2(g(x),x)dx + \iint_{D^\delta} v_x^2 dxdx + \int_0^\tau v v_x|_{x=r^\delta(t)} dt = 0,$$

where $g(x) := \sup\{t : (x,t) \in D^\delta\}$. Then, integrating this identity with respect to $\delta \in (0,\delta_0)$ and applying the last condition in (5.21), we get

$$\begin{aligned}
0 &\ge \int_0^{\delta_0} \left(\iint_{D^\delta} v_x^2 dxdt + \frac{1}{2}\int_0^\tau (v^2(t,R(t)+\delta))_x dt \right) d\delta \\
&= \int_0^{\delta_0} \left(\iint_{D^\delta} v_x^2 dxdt + \frac{1}{2}\int_0^\tau \frac{d}{d\delta}(v^2(t,R(t)+\delta))dt \right) d\delta \\
&= \int_0^{\delta_0} \iint_{D^\delta} v_x^2 dxdtd\delta + \frac{1}{2}\int_0^\tau v^2(t,R(t)+\delta)dt - \frac{1}{2}\lim_{\xi\to 0}\int_0^\tau v^2(t,R(t)+\xi)dt \\
&= \int_0^{\delta_0} \iint_{D^\delta} v_x^2 dxdxd\delta + \frac{1}{2}\int_0^\tau v^2(R(t)+\delta,t)dt.
\end{aligned}$$

This inequality immediately yields $v \equiv 0$ in Ω_τ for every $\tau \in (0,t_*)$ and completes the proof of Lemma 8. □

Therefore, $\theta^0 \equiv \theta$ in Ω_{t_*}. Together with (5.20), this identity also concludes the proof of Lemma 7. □

Let us continue the proof of Theorem 4. As we know from Theorem 2, a regular solution of the problem (A^0) is piecewise smooth with respect to t. Now we going to study the following question: On what does the size of every smoothness interval depend?

LEMMA 9. *Let us assume—under the conditions of Theorem 4—that for some $t', t'' \in (0,t_*)$, $t' < t''$ the following statements hold:*

$$|\theta_x(x,t')| < \text{const},\ x \in [0,1], \tag{5.22}$$

$$|\dot{R}(t)| < C(\delta),\ t \in [t', t''-\delta) \ \textit{for every } \delta > 0, \tag{5.23}$$

$$\limsup_{t\to t''-0} \dot{R}(t) = +\infty. \tag{5.24}$$

Then

$$\liminf_{t\to t''-0} r^-(t) = R(t''-0). \tag{5.25}$$

Proof of Lemma 9. Suppose that (5.25) does not hold. Note that the function θ is continuous inside of the domain $\{(x,t) : 0 < x < 1,\ t' < t < t''\}$. Therefore $r^-(t) > R(t)$ for every $t \in [t', t'')$. Then the converse of (5.25) implies that

$$r^-(t) - R(t) > \alpha > 0 \text{ for } t \in [t', t''']. \tag{5.26}$$

Because $\theta(R(t),t) = 0$, $\theta(r^-(t),t) = -L$, and $\theta(x,t') \geq -L$ for $x \in (R(t'), r^-(t'))$, we have $\theta(x,t) > -L$ for $t \in (t', t''']$, $R(t) < x < r^-(t)$. Therefore we can choose the positive value δ and the function $q \in C(t', t'')$ such that

$$q(t) - R(t) > \alpha/2, \quad t \in (t', t''), \tag{5.27}$$

$$\theta(q(t),t) > -L + \delta, \quad t \in (t', t''), \tag{5.28}$$

$$\theta(x,t') > -L + \delta, \quad x \in (R(t'), q(t')). \tag{5.29}$$

Making the transformation

$$\tau := t, \quad y := x - R(t)$$

and setting $S(\tau) := q(\tau) - R(\tau)$, we obtain that θ, R solve the following problem:

$$\theta_\tau - \dot{R}\theta_y - \theta_{yy} = 0 \text{ in } Q' \setminus \{y = 0\},$$
$$\theta_y(+0,\tau) - \theta_y(-0,\tau) = -L\dot{R},\ \theta(0,\tau) = 0, \quad \tau \in (t', t''),$$

where $Q' := \{(y,\tau) :\ \tau \in (t', t''), -R(\tau) < y < S(\tau)\}$.

We construct the following barrier function in the domain Q':

$$w(y) = \begin{cases} Ay, & y < 0, \\ (A - L\beta)(1 - e^{-\beta y})/\beta, & y > 0. \end{cases}$$

One can choose positive constants A, β so that the function $v = \theta - w$ is nonnegative on the parabolic boundary of the domain Q':

$$A = \max\{\max_{0 \leq x \leq 1} |\theta_x(x,t')|, \max_{t' \leq t \leq t''} |\theta^0(t)|\},$$

$$\frac{A - L\beta}{\beta}(1 - e^{-\beta y_1}) < \delta - L, \tag{5.30}$$

where

$$y_1 = \min\left\{\alpha/2, \frac{L - \delta}{\max_{0 \leq x \leq 1} |\theta_x(x,t')|}\right\}$$

and the constant α is taken from the inequality (5.27). We can choose β to satisfy (5.30), since the left-hand side of this inequality tends to $-L$ as $\beta \to +\infty$. Thus the function v satisfies the following problem:

$$v_\tau - \dot{R}v_y - v_{yy} = F(y,\tau) \text{ in } Q' \setminus \{y = 0\},$$
$$v(0,\tau) = 0,$$
$$v(y,\tau) \geq 0 \text{ on the parabolic boundary of the domain } Q',$$

where

$$F(y,\tau) = \begin{cases} -(\beta - \dot{R})w_y = (v_y(-0,\tau) - v_y(+0,\tau))w_y/L \text{ for } 0 < y < S(\tau), \\ \dot{R}(\tau)w_y \text{ for } -R(\tau) < y < 0. \end{cases}$$

Since $yw_y(y) \le 0$ and $\dot{R}(\tau) \ge 0$, the function $F(y,\tau)$ is nonnegative up to the time when $v(y,\tau) \ge 0$. However, under this condition the function v cannot have a negative minimum inside of the domain Q', i.e., $v(y,\tau) \ge 0$ in Q'. Hence

$$0 \le (v_y(+0,\tau) - v_y(-0,\tau))/L = \beta - \dot{R}(\tau), \quad \text{and}$$
$$\dot{R}(\tau) \le \beta, \ \tau \in (t', t'').$$

This contradicts the condition (5.24), and we have proved Lemma 9. □

Let us continue the proof of Theorem 4 by applying Theorem 2. One can suppose that (t_i, t^i) are maximal intervals from (3.13), i.e., for every $i \in I$

$$\text{either} \quad \limsup_{t \to t^i - 0} \dot{R}(t) = +\infty \quad \text{or } t^i = +\infty, \tag{5.31}$$

and one of the following three statements is valid:

(i) $t_i = 0$,

(ii) $t_i = t^k$ for some $k \in I$,

(iii) $t_i = \lim_{\{l\} \subset I} t^l$.

Let us take $i \in I$ such that $t_i > 0$. Assume that

$$\text{there exists } x_0 \in (R(t_i + 0), 1), \quad \text{such that } \theta(x_0, t_i) \le -L. \tag{5.32}$$

Then by the maximum principle there are sequences converging to $x_n \to x_0$, $\tau_n \to t_i - 0$ as $n \to \infty$ such that $\theta(x_n, \tau_n) < -L$. According to Lemma 7, we therefore have $r^+(\tau_n) > x_n$. The limit process as $n \to \infty$ gives us

$$\limsup_{t \to t_i - 0} r^+(t) \ge x_0 > R(t_i + 0). \tag{5.33}$$

Moreover, the assertion of Lemma 9 states

$$\liminf_{t \to t^k - 0} r^-(t) = R(t^k - 0) \quad \text{for every } k \in I, \ t^k < \infty.$$

Therefore the alternatives (ii) and (iii) imply

$$\liminf_{t \to t_i - 0} r^-(t) = R(t_i - 0) \le R(t_i + 0). \tag{5.34}$$

Since the function θ is continuous outside every neighborhood of the boundary $x = R(t)$, the conditions (5.33) and (5.34) and the simple connectedness of the set M immediately yield the inequality

$$\theta(x, t_i) \le -L, \ x \in (R(t_i + 0), x_0). \tag{5.35}$$

Let us choose $t' \in (t_i, t^i)$ such that $R(t') < x_0$ and

$$\theta(R(t'), t) < 0, \quad t \in (t_i, t^i). \tag{5.36}$$

Now we prove that the conditions (3.13), (5.35), (5.36) imply a contradiction, which means that the assumption (5.32) is not admissible. Using the estimate (3.13) we test the heat conduction equation for the function θ by $(x - R(t'))$ over the domain

$$\{(x,t) : t_i + \delta < t < t',\ R(t) < x < R(t')\},$$

$$(5.37)\quad -\int_{R(t_i+\delta)}^{R(t')} \theta(x, t_i+\delta)(x - R(t'))dx + \int_{t_i+\delta}^{t'} \theta_x(R(t)+0,t)(R(t) - R(t'))dt + \int_{t_i+\delta}^{t'} \theta(R(t'),t)dt = 0.$$

Due to the condition (5.36) the third integral is strictly negative. The second integral I_2^δ can be estimated with the help of the Stefan condition

$$\theta_x(R(t)+0,t) = \theta_x(R(t)-0,t) - L\dot{R}(t) \geq -L\dot{R}(t),$$

and hence

$$(5.38)\qquad I_2^\delta \leq \int_{t_i+\delta}^{t'} L\dot{R}(t)(R(t') - R(t))dt = \frac{L}{2}(R(t') - R(t_i+\delta))^2.$$

One can estimate the first integral I_1^δ in the identity (5.37) by using the inequality (5.35),

$$(5.39)\qquad \lim_{\delta\to 0} I_1^\delta = -\int_{R(t_i)}^{R(t')} \theta(x,t_i)(x - R(t'))dx \leq -\frac{L}{2}(R(t') - R(t_i))^2.$$

Therefore, letting $\delta \to 0$ in the identity (5.37), we obtain the contradiction

$$0 = \frac{L}{2}(R(t') - R(t_i))^2 - \frac{L}{2}(R(t') - R(t_i))^2 > 0.$$

So the converse of condition (5.32) is valid:

$$(5.40)\qquad \theta(x,t_i) > -L,\quad x \in (R(t_i+0), 1).$$

Since the set M is simply connected, (5.40) implies $M \cap \{t \geq t_i\} = \emptyset$. Then the statement of Lemma 9 yields $t^i = \infty$. Thus there exists at most one interval (t_i, t^i), $i \in I$ such that $t_i > 0$. If we denote $\bar{t} := t_i > 0$, then because of the properties (3.13) and (5.40), we obtain the validity of the inequalities (5.2) and (5.5) from the statement of Theorem 4.

In order to complete the proof of Theorem 4, we need only to prove the inequality (5.4). Note that, by the definition of the weak solution for the problem (A^0), the function U (see the formulation of the weak solution of the problem (A^0)) solves the heat conduction equation

$$U_t - U_{xx} = 0$$

in the domain

$$\{(x,t) : 0 < t < t_*,\ x \in (R(\bar{t}-0), R(\bar{t}+0))\}.$$

Here we have used the monotonicity of the function R on the interval $(0, t_*)$. Therefore

$$\lim_{t\to\bar{t}-0} U(x,t) = \lim_{t\to\bar{t}+0} U(x,t), \quad x \in (R(\bar{t}-0), R(\bar{t}+0)),$$

and

$$\theta(x,\bar{t}-0) + L = \theta(x,\bar{t}+0), \ x \in (R(\bar{t}-0), R(\bar{t}+0)). \tag{5.41}$$

From $\theta(x,t) \le 0$ for $x \in (0, R(t))$, $t \ge 0$, we deduce $\theta(x,\bar{t}+0) \le 0$, which together with (5.41) implies the inequality (5.4). The proof of Theorem 4 is now complete. □

COROLLARY 4. *If we replace the condition* (2.10) *by*

$$\theta^0(t) \le 0, \quad -L \le \theta^1(t) \le 0 \quad t \ge 0, \quad \theta_0(x) \le 0, \quad x \in (0,1), \tag{5.42}$$

then under the condition (5.1) *the statement of Theorem* 4 *also is valid.*

But the lifetime of the solution might be finite in this case, since the moving boundary can touch the fixed one, $x = 1$. This setup of the problem is simpler, because of $\theta^\varepsilon \le 0$, $\theta \le 0$ in Q_T and the functions R^ε, R are nondecreasing over the whole interval $(0,T)$. The conditions (5.42) also describe the one-phase initial-boundary data:

$$\theta^0(t) = 0, \ -L \le \theta^1(t) \le 0, \ t \ge 0;$$
$$\theta_0(x) = 0, \ x \in [0, R_0]; \ \theta_0(x) \le 0, \ x \in [R_0, 1].$$

It is obvious to see that after the jump of the free boundary, the solution becomes two-phase.

COROLLARY 5. *The results of Theorem* 4 *can be proved also for the problem* (A^0) *with Neuman boundary conditions*

$$\theta_x(i,t) = f^i(t), \quad 0 < t < T, \ i = 0,1,$$

where in addition to the conditions of Theorem 4 *we need the inequalities*

$$f^i(t) \ge 0, \quad 0 < t < T, \ i = 0,1.$$

The last condition makes sure that the temperature is nonpositive in the solid phase, and a new critical set $\{\theta < -L\}$ does not appear at the right boundary of the interval $x = 1$. However, similar to Corollary 4, we are able to study our solution only until the moving boundary touches the fixed one.

COROLLARY 6. *One can make the condition* (5.1) *less restrictive by introducing a finite number* N *of connected pieces of the set* $\{x : \theta_0(x) < -L\}$. *Then there exist at most* N *values* $t_i \in [0,T]$, $i = 1,\dots,M$, $M \le N$ *such that*

$$\sup_{t\in(t_i+\delta, t_{i+1}-\delta)} |\dot{R}(t)| \le C_i(\delta) \quad \textit{for every } \delta > 0, \ i = 0,\dots,M,$$
$$R(t_i+0) \ge R(t_i-0), \ i = 1,\dots,M-1,$$
$$\theta(x, t_i-0) \le -L, \ x \in (R(t_i-0), R(t_i+0)), \ i = 1,\dots,M,$$

where $t_0 = 0$, $t_{M+1} = T$.

6. Conclusion. We have studied the qualitative properties of the regular solutions of the problem (A^0). As follows from Theorem 4, the size of the jump of the free boundary is uniquely determined by the values of the temperature. Namely, the free boundary jumps exactly over the interval where $\theta \leq -L$. Please note that the arbitrary weak solution of the problem (A^0) does not have the previous property: The size of the jump can be anything. This fact implies, in particular, the nonuniqueness of the weak solution of the problem (A^0) (see, for example, [7]). Therefore the remaining open question is the uniqueness of a regular solution of the problem (A^0). It seems to be that the main difficulty of this task is to prove the uniqueness of a regular solution after the jump of the free boundary, i.e., with the following initial data at $t = \bar{t}$:

$$\begin{aligned}
&\theta(x, \bar{t}+0) = \vartheta(x), \ x \in (0,1), \\
&\vartheta \in C([0, R(\bar{t}+0)]), \ \vartheta(x) \leq 0, \ 0 \leq x \leq R(\bar{t}+0), \\
&\vartheta \in C([R(\bar{t}+0), 1]), \ \vartheta(x) \geq -L, \ R(\bar{t}+0) \leq x \leq 1, \\
&\lim_{x \to R(\bar{t}+0)+0} \vartheta(x) = -L, \lim_{x \to R(\bar{t}+0)-0} \theta(x) = 0.
\end{aligned}$$

The one-phase Stefan problem of this type was investigated by Fasano and Primicerio [2]. By using the maximum principle they proved the uniqueness of the solution. Unfortunately this method does not work for the two-phase problem.

REFERENCES

[1] B. SHERMAN, *A general one-phase Stefan problem,* Quart. Appl. Math., 28 (1970), pp. 377-382.

[2] A. FASANO AND M. PRIMICERIO, *A critical case for the solvability of Stefan-like problems,* Math. Meth. Appl. Sci., 5 (1983), pp. 84–96.

[3] A. FASANO, M. PRIMICERIO, S.D. HOWISON, J.R. OCKENDON, *Some remarks on the regularization of supercooled one-phase Stefan problems in one dimension,* Quart. Appl. Math., XLVIII (1990), pp. 153–168

[4] A. VISINTIN, *Stefan problem with a kinetic condition at the free boundary,* Ann. Mat. Pura Appl., 146 (1987), pp. 97–122.

[5] W. XIE, *The Stefan problem with a kinetic condition at the free boundary,* SIAM J. Math. Anal., 21 (1990), pp. 362–373.

[6] CH. CHARACH, B. ZALTZMAN, I. G. GÖTZ, *Interfacial kinetics effect in planar solidification problems without initial undercooling,* Math. Models Methods Appl. Sci., 4 (1994), pp. 331–354.

[7] A. M. MEIRMANOV, *The Stefan problem with surface tension in the three dimensional case with spherical symmetry: non-existence of the classical solution,* European J. Appl. Math., 5 (1994), pp. 1–19.

[8] H. KRÜGER, *Lineare parabolische Differentialgleichungen auf nichtzylindrischen Gebieten,* Inaugural-Dissertation, Augsburg 1984.

[9] A. M. MEIRMANOV, *The Stefan problem,* de Gruyter Exp. Math., Berlin, 1992.

[10] A. SARD, *The measure of the critical values of differentiable maps,* Bull. Amer. Math. Soc., 48 (1942), pp. 883–890.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 715–737, May 1995

011

NEW UNIQUENESS THEOREMS FOR THE ONE-DIMENSIONAL DRIFT-DIFFUSION SEMICONDUCTOR DEVICE EQUATIONS*

FATIHA ALABAU†

Abstract. We consider the one-dimensional drift-diffusion semiconductor device equations, under the assumption of zero generation-recombination. The uniqueness theorems that are given in the literature for this system do not hold for large values of the applied bias. The purpose of this paper is to introduce new techniques, which allow us to prove uniqueness theorems, in the case of symmetric p-n and p-i-n junctions, which are valid for arbitrary values of the applied bias. These techniques are essentially based on new monotonicity principles. As a direct consequence of these principles, we prove that the voltage-current characteristic of a symmetric p-n or p-i-n junction is strictly increasing.

Key words. nonlinear system, elliptic, semiconductor, uniqueness

AMS subject classifications. 35G30, 35J25, 35B50

1. Introduction.

In this paper we consider the uniqueness and the qualitative behavior of solutions of a system of differential equations that arises in the physics of semiconductor devices. This system, derived by Van Roosbroeck [16], forms a one-dimensional parameter-dependent system.

Under scaled form (see [11], [1]), the equations we consider are

$$\varepsilon\psi'' = n - p - N, \tag{1.1}$$

$$n' = n\psi' + J_n, \tag{1.2}$$

$$p' = -p\psi' - J_p, \tag{1.3}$$

in $\Omega =]-1, 1[$. This system is assumed to satisfy the boundary conditions

$$\psi(-1) = \psi_{-1}, \qquad \psi(1) = \psi_1, \tag{1.4}$$

$$n(-1) = n_{-1}, \qquad n(1) = n_1, \tag{1.5}$$

$$p(-1) = p_{-1}, \qquad p(1) = p_1, \tag{1.6}$$

where $\psi_{\pm 1}$, $n_{\pm 1}$, $p_{\pm 1}$ are given by

$$\psi_x = \log\left(\frac{N(x) + \sqrt{N^2(x) + 4\delta^4}}{2\delta^2}\right) - U(x), \qquad x = \pm 1, \tag{1.7}$$

$$n_x = \frac{N(x) + \sqrt{N^2(x) + 4\delta^4}}{2}, \qquad x = \pm 1, \tag{1.8}$$

* Received by the editors August 6, 1992; accepted for publication (in revised form) October 11, 1993.

† CeReMaB, Université Bordeaux I, 351, cours de la Libération, 33405 Talence Cedex, France.

$$p_x = \frac{-N(x) + \sqrt{N^2(x) + 4\delta^4}}{2}, \qquad x = \pm 1. \tag{1.9}$$

System (1.1)–(1.6) models the transport of electrons and holes in a semiconductor device under the effect of a parameter V, which acts on the device only through the boundary condition (1.4). Our model assumes that the mobilities are constant and that there are no generation-recombination effects. This last assumption (which is not physically realistic under strong forward bias) implies that the unknown electron and hole current densities are constant. We use this property throughout this paper, so that our results do not apply (at least directly) to the case of nonzero generation-recombination. The purpose of this paper is to derive qualitative new results on this simplified model, rather than to consider a more complex one. We shall discuss in §4 how these results can be extended to more complex models, in particular to the case of small generation-recombination terms. The unknowns are the functions ψ, n, p and the numbers J_n, J_p, which represent, respectively, the electrostatic potential, the electron and hole densities, and the electron and hole current densities. The given function N is called the doping profile. The number ε is a small positive constant, $-U(-1)$ and $-U(1)$ are the voltages applied at the endpoints of $\overline{\Omega}$, and δ is a positive number.

The parameter of interest in our study is

$$V = U(1) - U(-1); \tag{1.10}$$

this is called the applied bias and it will be allowed to take values in all $\mathbb{R}$, whereas ε and δ are kept fixed.

To express clearly that (1.1)–(1.6) is solved for a given V, we will refer to this system, when necessary, by the notation (1.1)–$(1.6)_V$.

Many results that concern the existence of solutions of the steady-state semiconductor device equations, (under a more general form than (1.1)–(1.6)), have been proved (see, e.g.,[13], [11], [8], [19]). In particular, for all strictly positive numbers ε and δ, for all V in $\mathbb{R}$, and for all piecewise-smooth functions N, (1.1)–$(1.6)_V$ admits at least one solution

$$(\psi, n, p, J_n, J_p) \in (H^1(\Omega))^3 \times \mathbb{R}^2. \tag{1.11}$$

Moreover, every solution (ψ, n, p, J_n, J_p) that satisfies (1.11) satisfies $n > 0$ and $p > 0$ on $\overline{\Omega}$.

The situation is wide open (see, e.g., [12], [18], [7]) as far as uniqueness or multiplicity of the solutions of the steady-state semiconductor device equations is concerned, even in the one-dimensional case. Moreover, the voltage-current curve (or characteristic) of the device that expresses the unknown current

$$I = J_n + J_p, \tag{1.12}$$

in terms of the applied bias, and that describes the electrical behavior of the semiconductor device, is of primary interest for the device engineers.

The purpose of our paper is to introduce new tools for considering the problem of uniqueness of solutions of (1.1)–$(1.6)_V$ for arbitrary values of V, and for studying the behavior of the voltage-current characteristic.

We first remark that, for physical reasons, the solution is not, in general, unique (see [12], [17], and [20] for numerical examples of multiplicity). Actually, the physical

performances of certain devices, like, for instance, thyristors, are explicitly based on the existence of multiple steady-state solutions. However, in multidimensional cases, uniqueness of solutions holds for sufficiently small $|V|$, uniformly with respect to the doping profile N (see [13], [11], and the references therein, [9], and see also the recent paper [14], which treats the case of weak solutions). Hence, it is important to distinguish between the case of small $|V|$ (that is close to equilibrium) for which uniqueness holds uniformly with respect to the doping profile N, and the case of nonsmall $|V|$, for which it is known, but not proved, that uniqueness does not hold uniformly with respect to the doping profile (see [18]). Therefore, uniqueness results for arbitrary V can be obtained only under additional assumptions on the doping profile N.

The methods used for proving uniqueness for small $|V|$, in the above-mentioned papers, and in [6] are essentially based on a perturbation argument for small $|V|$ (or small λ as in [6]), either based on a monotonicity property of the second member of the equation (1.1) with respect to ψ (when n and p are defined as functions of ψ through the continuity equations (1.2),(1.3) and the boundary conditions (1.5),(1.6)), or on a contraction property (as in [9]). This means that, for small $|V|$, the nonlinear system formed by the semiconductor equations is *weakly* coupled (see [12]). This is *no longer true* for arbitrary V. Therefore, one has to find other techniques, not only to get uniqueness results for arbitrary V, but also to find under which hypotheses on the doping profile N such uniqueness results hold.

One of the goals of this paper is to introduce new arguments for proving such uniqueness results. To formulate our results, we first recall here some basic definitions concerning semiconductor devices. The semiconductor is said to be *symmetric* when N is odd. A region where N is > 0 (resp., < 0) is called an *n-region* (resp., *p-region*). A region where $N = 0$ is called an *i-region*. The number of sign alterations of the doping profile N characterizes the type of the semiconductor. For a *pn-junction*, N has only one alteration of sign in the device. For a p-i-n junction, N has only one alteration of sign, but it is equal to zero in the i-region. The junction is said to be abrupt (resp., smooth) when N is discontinuous (resp., continuous) at the point where it changes its sign.

The purpose of this paper is threefold. First, we identify general a priori conditions on the solutions of (1.1)–(1.6) (see (2.2) and (2.3)), which are sufficient for proving global uniqueness theorems. This is done by introducing and proving some new monotonicity principles. More precisely, under the above-mentioned conditions on the solutions of (1.1)–(1.6), and when N is odd and satisfies additional hypotheses depending on the polarisation (i.e., on the sign of V) of the device, we show that the electric field $-\psi'$ satisfies a local and a global monotonicity property with respect to the current I. Whenever these monotonicity properties hold, we prove that (1.1)–(1.6) has a unique solution. We also give (in Remark 2.7), a physical interpretation of the above-mentioned conditions.

Second, we prove that these conditions are satisfied under additional assumptions on the doping profile N and for arbitrary large values of $|V|$. This allows us to prove that (1.1)–$(1.6)_V$ has a unique solution for all $V \leq 0$ (i.e., in the reverse-biased case) when N satisfies one of the following two hypotheses.

(i) N is odd and is in $\mathcal{C}^1(\overline{\Omega})$, $0 \leq N'$ on $\overline{\Omega}$ (which is an example of smooth pn or p-i-n junction);

(ii) N is odd, $N = 1$ on $[0, 1]$ (which is an example of abrupt pn-junction), and for all $V \geq V_0 = V_0(\delta, \varepsilon) > 0$ (i.e., in the forward-biased case) when (ii) holds.

The doping profiles N satisfying (i) or (ii) have at most one alteration of sign. Therefore our results partially answer the conjecture formulated by Mock [12] and by Rubinstein [17], which predicts that, when N has at most two alterations of sign in the device, then (1.1)–$(1.6)_V$ has a unique solution for all V.

Third, it is well known that numerical simulations and physical experiments predict that the voltage-current curve of a pn-junction is a strictly increasing function of V. We establish that this property is a direct consequence of the new monotonicity principles introduced in this paper and prove it for the devices described above. This is, to our knowledge, the first proof of a result which was known from numerical simulations, but was not demonstrated rigorously.

Furthermore, our results have obvious direct applications to current-driven models (see [10]), for which the current is assumed to be given, whereas the bias is an unknown.

The results of the present paper generalize our previous papers [4], [5] in the following sense. In [4], we consider two different types of devices, namely, bipolar membranes and unipolar (or single carrier) devices. In this last case only one type of charges (for instance, electrons) is transported through the device. Hence in the unipolar case, the system (1.1)–$(1.6)_V$ reduces to a system of two equations (derived from (1.1)–$(1.6)_V$, by setting one of the concentrations n or p equal to zero). For this case we prove global uniqueness of solutions for *arbitrary* V and for *arbitrary* doping profiles N. For the bipolar membrane case, one has to consider the full system (1.1)–(1.3). Of course, as is explained above, uniqueness of solutions of this full system for arbitrary V does not hold for arbitrary doping profiles N. In [4], we obtain uniqueness results in the case of bipolar membranes, for arbitrary large $|V|$, under the strong assumption of a *zero* doping profile N. The technique used in [4] is based on a decoupling method, which is valid whenever the doping profile N is piecewise constant. However, to obtain uniqueness results, we apply a maximum principle that for technical reasons is directly applicable only for a *zero* doping profile N. This argument is no longer applicable directly for a *nonzero* doping profile N. In [5], we obtained a first result for an example of *nonzero* odd doping profile (namely, $N(x) = sign(x)$). In [5] we use the decoupling method introduced in [4]. However, because of the technical difficulties generated by the presence of a *nonzero* doping profile, we need to apply a maximum principle on the first derivative of a function satisfying a true third-order differential equation (see Remark 2.6 in the present paper). As a consequence, we only prove in [5] a local uniqueness theorem for the solutions of the symmetrized version of system (1.1)–$(1.6)_V$ (which is the system derived from (1.1)-$(1.6)_V$ by only considering the symmetric solutions) when $V \leq 0$. In the present paper, we introduce a new decoupling method, which is valid even if the doping profile is not piecewise constant, but only in the case of symmetric (see Definition 1.2 in this paper) solutions. As a consequence of this new decoupling, we prove more general results, and in particular that (1.1)–$(1.6)_V$ has a unique solution (which is, of course, necessarily symmetric) under the above given assumptions on N and V. We also prove the monotonicity of the voltage-current curve in the cases given above.

Note here that the present results do not apply to the bipolar membrane case, which is under consideration in [4]. In this last case there is no symmetry assumption because of the boundary conditions for the bipolar membranes, which are slightly different from the boundary conditions for semiconductors.

Note also that the local monotonicity property mentioned in this paper has indeed been introduced in an informal way in [5].

The paper is organized as follows. Section 2 is devoted to the reverse biased case. In §2.1 we derive, for general symmetric devices, sufficient a priori conditions on the solutions for proving uniqueness and monotonicity of the current. In §2.2 we give applications of these results to symmetric smooth and abrupt pn-junctions. In §3 we study the forward-biased case. In §3.1 we give a priori sufficient conditions on the solutions for proving uniqueness of solutions and monotonicity of the current in the case of abrupt pn-junctions. In §3.2 we prove that these a priori conditions are satisfied for strong forward biases.

Throughout this paper solutions of (1.1)–(1.6)$_V$ will always be assumed to satisfy (1.11).

We assume from now on that the semiconductor device is symmetric (see [4] and [3] for nonsymmetric devices). This assumption gives the following important remarks.

Remark 1.1. Let (ψ, n, p, J_n, J_p) be a solution of (1.1)–(1.6). We define the corresponding symmetrized vector $(\psi^s, n^s, p^s, J_n^s, J_p^s)$ in $(H^1(\Omega))^3 \times \mathbb{R}^2$ by

$$\psi^s(x) = -\psi(-x),\ n^s(x) = p(-x),\ p^s(x) = n(-x) \quad \forall x \text{ in } \overline{\Omega}, \tag{1.13}$$

and

$$J_n^s = J_p,\ J_p^s = J_n. \tag{1.14}$$

Then it is easy to check that $(\psi^s, n^s, p^s, J_n^s, J_p^s)$ is also a solution of (1.1)–(1.6).

DEFINITION 1.2. *A solution of* (1.1)–(1.6) *is said to be* symmetric *if it coincides with its symmetrized vector.*

Remark 1.3. We deduce from Remark 1.1 that if (1.1)–(1.6) has a nonsymmetric solution then it has multiple solutions.

2. Reverse-biased symmetric semiconductor devices. A symmetric semiconductor is said to be *reverse-biased* when $N(1)V < 0$. It is at *equilibrium* when $V = 0$. To fix the sign of V in the reverse-biased case, we assume, without loss of generality, within this section that

$$0 < N(1). \tag{2.1}$$

We set $\Omega_+ = (0, 1)$.

2.1. Sufficient conditions for proving uniqueness of solutions and monotonicity of the current.

THEOREM 2.1. *Assume that* $N|_{\overline{\Omega}_+} \in C^1(\overline{\Omega}_+)$ *and let* $V \leq 0$ *be such that the following two hypotheses are verified.*

(i) *All solutions* (ψ, n, p, J_n, J_p) *of* (1.1)–(1.6)$_V$ *are such that*

$$I\psi'(x) \leq 0 \quad \forall x \in \overline{\Omega}_+, \tag{2.2}$$

where I *is given by* (1.12).

(ii) *There exists a symmetric solution* (ψ, n, p, J_n, J_p) *of* (1.1)–(1.6)$_V$ *that is such that*

$$0 \leq (n - p)(x) \quad \forall x \in \overline{\Omega}_+. \tag{2.3}$$

Then (1.1)-(1.6)$_V$ *has a unique solution satisfying* (1.11).

The proof is divided in two steps. We first prove that under hypothesis (i) of Theorem 2.1 all solutions of (1.1)–(1.6)$_V$ are symmetric. We then prove uniqueness in the class of symmetric solutions. These two results are independent, so that we give them separately. Let us first give the result of symmetry.

THEOREM 2.2. *Assume that* $N|_{\overline{\Omega}_+} \in \mathcal{C}^1(\Omega_+)$ *and let* $V \leq 0$ *be such that there exists a solution* $(\bar{\psi}, \bar{n}, \bar{p}, \bar{J}_n, \bar{J}_p)$ *of* (1.1)-(1.6)$_V$ *satisfying* (2.2). *Then this solution is symmetric.*

Proof. We set

$$I = \bar{J}_n + \bar{J}_p, \ J = \bar{J}_n - \bar{J}_p,$$

$$\psi(x) = \bar{\psi}(x), \ n(x) = \bar{n}(x), \ p(x) = \bar{p}(x) \quad \forall x \in \overline{\Omega}_+,$$

and

$$\phi(x) = -\bar{\psi}(-x), \ \tilde{n}(x) = \bar{p}(-x), \ \tilde{p}(x) = \bar{n}(-x) \quad \forall x \in \overline{\Omega}_+.$$

Then $(\psi, n, p, \bar{J}_n, \bar{J}_p)$ and $(\phi, \tilde{n}, \tilde{p}, \bar{J}_p, \bar{J}_n)$ are both solutions of (1.1)–(1.3) on Ω_+.

Because of Remark 1.1, we can assume without loss of generality that

$$J \leq 0. \tag{2.4}$$

Let us first assume that $V < 0$. Since (2.1) holds, we have

$$I < 0. \tag{2.5}$$

From (1.1)–(1.3), we obtain

$$4(np - \tilde{n}\tilde{p})' = -2I\varepsilon(\psi - \phi)'' + 2J(n + p + \tilde{n} + \tilde{p}) \quad \text{on } \Omega_+.$$

From (2.4) and since $(\bar{\psi}, \bar{n}, \bar{p})$ is in $\mathcal{C}^1(\overline{\Omega}) \times (\mathcal{C}(\overline{\Omega}))^2$, we deduce that

$$4(np - \tilde{n}\tilde{p}) \leq -2I\varepsilon(\psi - \phi)' \quad \text{on } \Omega_+. \tag{2.6}$$

Since $(\psi, n, p, \bar{J}_n, \bar{J}_p)$ and $(\phi, \tilde{n}, \tilde{p}, \bar{J}_p, \bar{J}_n)$ are solutions of (1.1)–(1.3), we obtain

$$\varepsilon(\psi - \phi)^{(3)} = (n + p)\psi' - (\tilde{n} + \tilde{p})\phi'.$$

We now replace in this last equation $n + p$ by $\left((n - p)^2 + 4np\right)^{1/2}$, do the same thing for $\tilde{n} + \tilde{p}$, and use (2.2) and (2.5), (2.6) to obtain

$$\varepsilon(\psi - \phi)^{(3)} \leq (\psi - \phi)''a_1 + (\psi - \phi)'a_0 \quad \text{on } \Omega_+, \tag{2.7}$$

where a_1 is a smooth function on $\overline{\Omega}_+$ and where a_0 is given by

$$a_0 = \tilde{n} + \tilde{p} - \frac{2I\varepsilon\psi'}{n + p + \tilde{n} + \tilde{p}}.$$

From (2.2), it follows that

$$a_0 > 0 \text{ on } \overline{\Omega}_+.$$

Since (2.6) holds, we have

$$0 \leq (\psi - \phi)'(1).$$

However, we also have $(\psi - \phi)'(0) = 0$, so that the maximum principle (see [15]) applied to (2.7) implies

$$0 \leq (\psi - \phi)' \quad \text{on } \Omega_+. \tag{2.8}$$

Since

$$((n - \tilde{n}) \exp(-\psi))' = \tilde{n} \exp(-\psi)(\psi - \phi)' + J \exp(-\psi),$$

we obtain

$$(n - \tilde{n})(x) \leq -J \exp(\psi(x)) \int_x^1 \exp(-\psi(t))dt \quad \forall x \text{ in } \Omega_+.$$

From (2.2) and (2.4)-(2.5), we deduce that

$$(n - \tilde{n})(x) \leq -J(1 - x) \quad \forall x \in \Omega_+.$$

By following a similar argument for $(p - \tilde{p})$, we obtain

$$(\psi - \phi)'' \leq 0 \text{ in } \Omega_+.$$

This together with (2.8) implies that $\psi = \phi$ on $\overline{\Omega}_+$. The equations $n = \tilde{n}$, $p = \tilde{p}$, and $J = 0$ follow at once. This proves the theorem when $V < 0$. If $V = 0$, we remark that inequality (2.7) becomes an equality, so that the same conclusion holds.

This concludes the proof.

We now prove uniqueness in the class of symmetric solutions under weaker hypotheses than those of Theorem 2.1.

THEOREM 2.3. *Assume that* $N|_{\overline{\Omega}_+} \in C^1(\overline{\Omega}_+)$ *and let* $V \leq 0$ *be such that there exists a symmetric solution* (ψ, n, p, J_n, J_p) *of* (1.1)–(1.6)$_V$ *satisfying* (2.2) *and* (2.3). *Then, the problem* (1.1)–(1.6)$_V$ *has a unique symmetric solution satisfying* (1.11).

Proof. Assume first that $V < 0$ and let $(\phi, \tilde{n}, \tilde{p}, \tilde{J}_n, \tilde{J}_p)$ be another symmetric solution of (1.1)-(1.6)$_V$.

We set

$$I = J_n + J_p \ , \ \tilde{I} = \tilde{J}_n + \tilde{J}_p$$

and

$$J = J_n - J_p \ , \ \tilde{J} = \tilde{J}_n - \tilde{J}_p.$$

As in the proof of Theorem 2.2, we replace $n + p$ by $\left((n - p)^2 + 4np\right)^{1/2}$, so that, finally, we obtain

$$\begin{aligned} \varepsilon(\psi - \phi)^{(3)} = (\psi - \phi)'' b_1 + \frac{4\psi'}{n + p + \tilde{n} + \tilde{p}}(np - \tilde{n}\tilde{p}) \\ + (\tilde{n} + \tilde{p})(\psi - \phi)' + I - \tilde{I} \text{ on } \Omega_+, \end{aligned} \tag{2.9}$$

where b_1 is a smooth function on $\overline{\Omega}_+$.

Since $J = \tilde{J} = 0$, we have

$$4(np - \tilde{n}\tilde{p})' = -2I(n-p) + 2\tilde{I}(\tilde{n} - \tilde{p}) \quad \text{on } \Omega_+. \tag{2.10}$$

Let M be a primitive of N and x_0 be an arbitrary point of $\overline{\Omega}_+$. Integration of (2.10) yields

$$\begin{aligned} 4(np - \tilde{n}\tilde{p})(x) = &-2(I - \tilde{I})\,(\varepsilon\psi'(x) + M(x) - (\varepsilon\psi'(x_0) + M(x_0))) \\ &-2\tilde{I}\varepsilon(\psi - \phi)'(x) + 2\tilde{I}\varepsilon(\psi - \phi)'(x_0) \\ &+4(np - \tilde{n}\tilde{p})(x_0) \quad \forall x \in \overline{\Omega}_+. \end{aligned} \tag{2.11}$$

We use this last equality in (2.9) to obtain

$$\begin{aligned} \varepsilon(\psi - \phi)^{(3)}(x) = &(\psi - \phi)''(x)b_1(x) + (\psi - \phi)'(x)b_0(x) \\ &+(I - \tilde{I})(1 - \frac{2\psi'(x)}{(n + p + \tilde{n} + \tilde{p})(x)}\,(\varepsilon\psi'(x) + M(x) \\ &- (\varepsilon\psi'(x_0) + M(x_0)))) \\ &+\frac{\psi'(x)}{(n + p + \tilde{n} + \tilde{p})(x)}\left(2\tilde{I}\varepsilon(\psi - \phi)'(x_0) + 4(np - \tilde{n}\tilde{p})(x_0)\right) \quad \forall x \in \overline{\Omega}_+, \end{aligned} \tag{2.12}$$

where b_0 is given by

$$b_0 = \tilde{n} + \tilde{p} - \frac{2\tilde{I}\varepsilon\psi'}{n + p + \tilde{n} + \tilde{p}}\,.$$

Using (2.2) and the fact that I and $\tilde{I}$ have the same sign, we obtain

$$0 < b_0 \text{ on } \overline{\Omega}_+.$$

Note that we can assume, without loss of generality, that

$$I - \tilde{I} \leq 0. \tag{2.13}$$

We claim that the following inequalities:

$$0 \leq (\psi - \phi)'(0) \quad \text{and} \quad 0 \leq (\psi - \phi)'(1) \tag{2.14}$$

hold. To prove this, we proceed as in [5, Lem. 3.3]. Assume to the contrary that we have

$$0 > (\psi - \phi)'(0) \quad \text{or} \quad 0 > (\psi - \phi)'(1).$$

We set $E = \{x \in [0,1], (\psi - \phi)'(x) = 0\}$.

Let us first assume that $(\psi - \phi)'(1) < 0$ holds and let $y^\star$ be the largest element of E. Hence, we have

$$(\psi - \phi)' \leq 0 \quad \text{on } [y^\star, 1] \quad \text{and} \quad (\psi - \phi)'(y^\star) = 0.$$

From (2.13) and

$$((n - \tilde{n})\exp(-\psi))' = \tilde{n}\exp(-\psi)(\psi - \phi)' + ((I - \tilde{I})/2)\exp(-\psi)\,, \tag{2.15}$$

we deduce that

$$0 \leq n - \tilde{n} \quad \text{on } [y^\star, 1].$$

In a similar way, we obtain

$$0 \leq -(p - \tilde{p}) \quad \text{on } [y^\star, 1].$$

Now, however, (1.1) implies

$$\varepsilon(\psi - \phi)'' = (n - \tilde{n}) - (p - \tilde{p}). \tag{2.16}$$

Hence, we obtain

$$0 \leq (\psi - \phi)'' \quad \text{on } [y^\star, 1].$$

This contradicts the inequality $(\psi - \phi)'(1) < 0$. Therefore, we can now assume that we have

$$(\psi - \phi)'(0) < 0 \quad \text{and} \quad 0 \leq (\psi - \phi)'(1). \tag{2.17}$$

Let $x^\star$ be the smallest element of E. Hence, we have

$$(\psi - \phi)' \leq 0 \quad \text{on } [0, x^\star] \quad \text{and} \quad (\psi - \phi)'(x^\star) = 0.$$

From (2.15) and since $\psi(0) = 0$, we deduce that

$$(n - \tilde{n})(x) \leq (n - \tilde{n})(0)\exp(\psi(x)) \quad \forall x \in [0, x^\star].$$

In a similar way, we obtain

$$-(p - \tilde{p})(x) \leq -(p - \tilde{p})(0)\exp(-\psi(x)) \quad \forall x \in [0, x^\star].$$

From (2.16) and since

$$(n - p)(0) = (\tilde{n} - \tilde{p})(0) = 0 \tag{2.18}$$

holds, we obtain

$$\varepsilon(\psi - \phi)''(x) \leq 2(n - \tilde{n})(0)\sinh(\psi(x)) \quad \forall x \in [0, x^\star]. \tag{2.19}$$

We now set $x = 0$ and $x_0 = 1$ in (2.11) and use (2.18) to derive

$$\begin{aligned} 4(n - \tilde{n})(0)(n + \tilde{n})(0) = {} & 2(I - \tilde{I})\left(\varepsilon\psi'(1) + M(1) - (\varepsilon\psi'(0) + M(0))\right) \\ & -2\tilde{I}\varepsilon(\psi - \phi)'(0) + 2\tilde{I}\varepsilon(\psi - \phi)'(1). \end{aligned} \tag{2.20}$$

From property (2.3) we deduce that the function $\varepsilon\psi' + M$ is increasing on $\overline{\Omega}_+$. Since (2.5) (with I replaced by $\tilde{I}$), (2.13), and (2.17) hold we deduce that

$$(n - \tilde{n})(0) \leq 0. \tag{2.21}$$

Now from (2.5), (2.2) and $\psi(0) = 0$ we deduce that

$$0 \leq \sinh(\psi(x)) \quad \forall x \in [0, 1].$$

Using this last inequality and (2.21) in (2.19), we obtain

$$(\psi - \phi)'' \leq 0 \text{ on } [0, x^*],$$

which contradicts (2.17).

Hence we proved inequality (2.14).

The function $\varepsilon\psi' + M$ is increasing on $\overline{\Omega}_+$. Thus, from the inequalities $b_0 > 0$, (2.2), (2.5), (2.13)–(2.14), and from the maximum principle applied to (2.12), we conclude that

$$0 \leq (\psi - \phi)' \text{ on } \overline{\Omega}_+. \tag{2.22}$$

This proves the theorem when $V < 0$. Moreover, if $V = 0$, then

$$I = \tilde{I}, \qquad np = \tilde{n}\tilde{p} = n_1 p_1 \quad \text{on} \quad \Omega,$$

so that we get to the same conclusion. Hence theorem 2.3 is proved.

Observe that Theorem 2.1 follows from Theorems 2.2 and 2.3.

Remark 2.4. We have proved and applied in the proof of Theorem 2.3 two new monotonicity principles that are satisfied by the symmetric solutions of (1.1)–$(1.6)_V$ in the reverse-biased case. Let ψ and ϕ be the electrostatic potentials, I and $\tilde{I}$ be the two respective currents corresponding to two symmetric solutions. Then the first monotonicity principle expresses a local monotonicity property of the electric field with respect to the current in the following sense.

If (2.2), (2.3) hold, then

$$I \leq \tilde{I} \implies 0 \leq (\psi - \phi)'(0) \quad \text{and} \quad 0 \leq (\psi - \phi)'(1).$$

The second monotonicity principle expresses a global monotonicity property of the electric field with respect to the current in the following sense.

If (2.2), (2.3) hold, then

$$I \leq \tilde{I} \quad \text{and} \quad 0 \leq (\psi - \phi)'(0), \qquad 0 \leq (\psi - \phi)'(1) \implies 0 \leq (\psi - \phi)' \quad \text{on } \overline{\Omega}_+.$$

Remark 2.5. It is important to remark that the proof of the global monotonicity property mentioned in Remark 2.4 relies on a new decoupling method (compared to the one introduced in [4] and [3]), which uses the properties of the symmetric solutions of (1.1)–(1.6) and allows us to prove that $(\psi - \phi)'$ satisfies a second-order differential equation, namely (2.12), which is decoupled from the equations satisfied by $n - \tilde{n}$ and $p - \tilde{p}$.

Remark 2.6. When the doping profile is not necessarily odd but is piecewise-constant, we have introduced in [3] a general decoupling method. One main difficulty in this case is to justify a maximum principle on the derivative of the solution of a third-order differential equation, for which the zero-order coefficient is not vanishing. Our new decoupling method allows us to avoid this difficulty in the case of symmetric devices.

Remark 2.7. The a priori properties (2.2) and (2.3) have a precise physical meaning. The inequality (2.2) means that the current I and the electric field $-\psi'$ flow in the same direction. The inequality (2.3) means that the electron density is larger than the hole density. We will see in §2.2 that this property is always satisfied when N has only one alteration of sign. Moreover, we can remark that we have already used (2.2) in [5] to justify a generalized maximum principle (see [15]), which

allowed us to prove that the symmetric solutions of (1.1)–(1.6) are locally unique (see also [3] for the forward-biased case). The new decoupling technique introduced here and (2.3) are fundamental for the proof of global uniqueness.

We now study the monotonicity of the voltage-current curve.

THEOREM 2.8. *Assume that* $N|_{\overline{\Omega}_+} \in C^1(\overline{\Omega}_+)$ *and let* V *and* $\tilde{V}$ *be such that* $V < \tilde{V} \leq 0$. *Assume moreover that there exists a symmetric solution* (ψ, n, p, J_n, J_p) *(resp.,* $(\phi, \tilde{n}, \tilde{p}, \tilde{J}_n, \tilde{J}_p)$*) of* (1.1)–$(1.6)_V$ *(resp.,* (1.1)–$(1.6)_{\tilde{V}}$*) satisfying* (2.2) *and* (2.3).

Then the following property holds:

$$I < \tilde{I}, \tag{2.23}$$

where $I = J_n + J_p$ *and* $\tilde{I} = \tilde{J}_n + \tilde{J}_p$.

Proof. From the hypotheses and Theorem 2.3, we know that (1.1)–$(1.6)_V$ (resp., (1.1)–$(1.6)_{\tilde{V}}$) has a unique symmetric solution. We claim that (2.23) holds. To the contrary, assume that

$$0 \leq I - \tilde{I}. \tag{2.24}$$

As in the proof of Theorem 2.3, we show that if $(\psi - \phi)'(0) \leq 0$ and $(\psi - \phi)'(1) \leq 0$ hold, then we have $(\psi - \phi)' \leq 0$ on $\overline{\Omega}_+$, which contradicts the hypothesis $V < \tilde{V}$. Hence we have either $0 < (\psi - \phi)'(0)$ or $0 < (\psi - \phi)'(1)$.

Let us first assume that $0 < (\psi - \phi)'(1)$. Then, we claim that $(\psi - \phi)'$ changes sign on $\overline{\Omega}_+$. Assume to the contrary that $(\psi - \phi)'$ keeps a constant sign on $\overline{\Omega}_+$. We deduce then from (2.24) that

$$0 < (\psi - \phi)' \quad \text{on } \overline{\Omega}_+.$$

From this, we get

$$2(I - \tilde{I}) \cosh \psi < ((n - \tilde{n}) \exp(-\psi) - (p - \tilde{p}) \exp(\psi))' \quad \text{on } \overline{\Omega}_+,$$

which contradicts (2.24).

Therefore $(\psi - \phi)'$ changes sign on $\overline{\Omega}_+$. We now proceed as in the proof of Theorem 2.3. Let $y^\star$ be the largest element of

$$X = \{x \in \overline{\Omega}_+, (\psi - \phi)'(x) = 0\}.$$

We use the inequalities $0 \leq (\psi - \phi)'$ on $[y^\star, 1]$ and (2.24) in (2.15) to deduce that

$$(n - \tilde{n}) \leq 0 \quad \text{on } [y^\star, 1].$$

In a similar way, we obtain

$$-(p - \tilde{p}) \leq 0 \quad \text{on } [y^\star, 1].$$

Thus we have $(\psi - \phi)'' \leq 0$ on $[y^\star, 1]$, which contradicts $0 < (\psi - \phi)'(1)$. Hence, we proved that we have

$$(\psi - \phi)'(1) \leq 0.$$

Assume now that $0 < (\psi - \phi)'(0)$, and let $x^\star$ be the smallest element of X. Since we have $0 \leq (\psi - \phi)'$ on $[0, x^\star]$, we deduce that

$$2(n - \tilde{n})(0) \sinh \psi(x) \leq \varepsilon(\psi - \phi)''(x) \quad \forall x \in [0, x^\star].$$

Now setting $x = 0$ and $x_0 = 1$ in (2.11), we obtain

$$0 \leq (n - \tilde{n})(0).$$

From (2.2) and (2.5), we obtain $0 \leq \psi$ on $\overline{\Omega}_+$. So that finally, we obtain

$$0 \leq (\psi - \phi)'' \quad \text{on } [0, x^\star],$$

which contradicts the inequality $0 < (\psi - \phi)'(0)$. Hence (2.24) is impossible, and this concludes the proof.

2.2. Applications to reverse-biased symmetric pn-junctions. We gave in the previous subsection sufficient a priori conditions, in the reverse-biased case, to be satisfied by the solutions of (1.1)–(1.6) in order to prove uniqueness and monotonicity of the current. We now give applications of these results.

THEOREM 2.9. *Assume that* $N|_{\overline{\Omega}} \in \mathcal{C}^1(\overline{\Omega})$ *and that*

$$0 \leq N' \quad \text{on } \overline{\Omega}_+.$$

Then for every $V \leq 0$, *(1.1)-(1.6)*$_V$ *has a unique solution. Moreover, the function*

$$V \in]-\infty, 0] \mapsto I \in \mathbb{R}$$

is continuous and strictly increasing.

Proof. Assume that $V \leq 0$ and let (ψ, n, p, J_n, J_p) be a solution of (1.1)–(1.6)$_V$. Since N is in $\mathcal{C}^1(\overline{\Omega})$, we obtain

$$\varepsilon\psi^{(3)} = (n + p)\psi' + I - N' \quad \text{on } \overline{\Omega}. \tag{2.25}$$

Moreover, we also have

$$\psi''(-1) = 0, \qquad \psi''(1) = 0. \tag{2.26}$$

Since $N(1)V \leq 0$, we obtain $I \leq 0$. The maximum principle applied to (2.25) subject to the Neumann boundary conditions (2.26) implies

$$0 \leq \psi' \quad \text{on } \overline{\Omega}_+.$$

Hence (2.2) is satisfied for every $V \leq 0$ and for every solution of (1.1)–(1.6)$_V$. We deduce from Theorem 2.2 that for every $V \leq 0$, all the solutions of (1.1)–(1.6)$_V$ are symmetric. Hence, we have

$$\varepsilon(n - p)'' = (n - p)(\varepsilon\psi'^2 + n + p) - N(n + p) \quad \text{on } \overline{\Omega}_+, \tag{2.27}$$

and

$$(n - p)(0) = 0, \qquad (n - p)(1) = N(1) \geq 0.$$

Since $0 \leq N$ on $\overline{\Omega}_+$, an easy application of the maximum principle to (2.27) proves that $n - p$ satisfies (2.3).

Theorems 2.3 and 2.8 prove that (1.1)–(1.6)$_V$ has a unique solution for every $V \leq 0$ and that the current is a strictly increasing function of V on $(-\infty, 0]$. We deduce the continuity of the voltage-current curve from this uniqueness result and from [11, Thm. 3.5.2].

Remark 2.10. In the case of abrupt symmetric pn-junction, N is discontinuous at 0, so that ψ'' is also discontinuous at 0. Hence (2.25) does not hold in all Ω and information is lost at the junction $x = 0$. This is the reason why it is not possible, in this case, to proceed as in Theorem 2.9 for proving (2.2). However, using a result of [6] on the qualitative properties of the solutions of (1.1)–(1.6) in the reverse-biased case, we can prove the following theorem.

THEOREM 2.11. *Assume that* $N|_{\overline{\Omega}_+} = 1$. *Then for every* $V \leq 0$, (1.1)–$(1.6)_V$ *has a unique solution. Moreover, the function*

$$V \in]-\infty, 0] \mapsto I \in \mathbb{R}$$

is continuous and strictly increasing.

Proof. Since $N(1)V \leq 0$, we have $I \leq 0$. This together with [6, Prop. 10] prove (2.2). Moreover, as in the proof of Theorem 2.9, (2.3) holds for every symmetric solution. This concludes the proof.

Remark 2.12. Mock [13] and Rubinstein [17] made the conjecture that (1.1)–$(1.6)_V$ can have multiple solutions for certain values of V (indeed in the forward-biased case) only when N has at least three alterations of sign in the device. As a "corollary" of this conjecture, uniqueness should hold for pn-junctions. Theorems 2.9 and 2.11 partially answer this conjecture. Moreover the fact that N has only one alteration of sign in the above-mentioned theorems is essential for proving inequality (2.3).

Remark 2.13. In addition to Remark 2.7, we can note that inequality (2.3) expresses, in the case of pn-junctions, the following physical property: in an n-type region of the device (that is, in a region where $N > 0$) the electron density is larger than the hole density.

3. Forward-biased symmetric abrupt pn-junctions. A symmetric semiconductor is said to be *forward-biased* when $0 < N(1)V$.

It is well known that the semiconductor behaves very differently in reverse and forward-biased cases. We will see now that this last case generates several difficulties. First, it is easy to check that property (2.2) does not hold for all $V > 0$. On the other hand, we will see in this section that even when (2.2) holds, the proofs of theorems similar to Theorems 2.3, 2.8, and 2.11 require important changes, due to the fact that the electric field has an opposite direction compared to the one it has in the reverse-biased case.

3.1. Sufficient conditions for proving uniqueness of the solutions and monotonicity of the current.

THEOREM 3.1. *Assume that*

$$N|_{\overline{\Omega}_+} = 1, \tag{3.1}$$

and let $V > 0$ *be such that hypothesis* (i) *of Theorem* 2.1 *holds. Then* (1.1)–$(1.6)_V$ *has a unique solution.*

We first prove, as in the reverse-biased case, the following result of symmetry, whose proof (similar to that of Theorem 2.2) is left to the reader.

THEOREM 3.2. *Assume that* $N|_{\overline{\Omega}_+} \in \mathcal{C}^1(\overline{\Omega}_+)$ *and let* $V > 0$ *be such that there exists a solution* $(\bar{\psi}, \bar{n}, \bar{p}, \bar{J}_n, \bar{J}_p)$ *of* (1.1)–$(1.6)_V$ *satisfying* (2.2). *Then this solution is symmetric.*

We now prove uniqueness in the class of symmetric solutions under a stronger assumption on N.

THEOREM 3.3. *Assume that* (3.1) *holds and let* $V > 0$ *be such that every symmetric solution of* (1.1)–(1.6)$_V$ *satisfies* (2.2). *Then* (1.1)–(1.6)$_V$ *has a unique symmetric solution.*

Proof. Let (ψ, n, p, J_n, J_p) and $(\phi, \tilde{n}, \tilde{p}, \tilde{J}_n, \tilde{J}_p)$ be two symmetric solutions of (1.1)-(1.6)$_V$. We set

$$I = J_n + J_p, \qquad \tilde{I} = \tilde{J}_n + \tilde{J}_p.$$

We can assume without loss of generality that (2.13) holds. As for the reverse-biased case in Theorem 2.3, this implies that

$$0 \le (\psi - \phi)'(0) \quad \text{and} \quad 0 \le (\psi - \phi)'(1). \tag{3.2}$$

We claim that

$$0 \le (\psi - \phi)' \quad \text{on } \overline{\Omega}_+.$$

Assume to the contrary that $(\psi - \phi)'$ has a strictly negative minimum at y_0 in Ω_+. Equation (2.12) is still satisfied, but $M(x)$ is now equal to x for all x in $\overline{\Omega}_+$ (since (3.1) holds). Moreover, from (2.2) we deduce that $b_0 > 0$. We now choose $x_0 = y_0$ in (2.12). We deduce from the maximum principle applied to (2.12) that

$$2\tilde{I}\varepsilon(\psi - \phi)'(y_0) + 4(np - \tilde{n}\tilde{p})(y_0) < 0. \tag{3.3}$$

On the other hand, there exists two unknown constants α and $\tilde{\alpha}$ such that

$$n + p = \varepsilon\frac{\psi'^2}{2} + \psi + \alpha \quad \text{and} \quad \tilde{n} + \tilde{p} = \varepsilon\frac{\phi'^2}{2} + \phi + \tilde{\alpha} \quad \text{on } \overline{\Omega}_+.$$

Hence we have, since hypothesis (i) of theorem 2.1, $0 < I$ and (3.2) hold,

$$0 \le \alpha - \tilde{\alpha} = -(\varepsilon/2)(\psi - \phi)'(1)(\psi + \phi)'(1). \tag{3.4}$$

Since (3.1) holds, any symmetric solution of (1.1)–(1.6)$_V$ satisfies (2.3). So we can replace $n - p$ by $\left((n+p)^2 - 4np\right)^{1/2}$. We do the same thing for $\tilde{n} - \tilde{p}$; this leads to

$$\begin{aligned} &\varepsilon(n - p + \tilde{n} - \tilde{p})(\psi - \phi)''(x) \\ &= (\psi - \phi)'(x)c_1(x) + (n + p + \tilde{n} + \tilde{p})(x)\left((\psi - \phi)(x) + \alpha - \tilde{\alpha}\right) \\ &\quad + 2(I - \tilde{I})\left(\varepsilon\psi'(x) + x - (\varepsilon\psi'(y_0) + y_0)\right) \\ &\quad - \left(2\tilde{I}\varepsilon(\psi - \phi)'(y_0) + 4(np - \tilde{n}\tilde{p})(y_0)\right) \quad \text{on } \overline{\Omega}_+, \end{aligned} \tag{3.5}$$

where c_1 is a smooth function on $\overline{\Omega}_+$.

Now assume first that

$$0 < (\psi - \phi)'(0).$$

Then $(\psi - \phi)$ attains a strictly positive maximum at a point $x^\star \in (0, y_0)$. However, we obtain a contradiction by applying the maximum principle to (3.5) and using (2.3) and (3.3)–(3.4). Hence we proved that

$$(\psi - \phi)'(0) = 0,$$

which implies

$$0 \leq (n - \tilde{n})(0).$$

We now choose $x_0 = 0$ in (2.12) and apply the maximum principle once again to this equation. This leads to

$$0 \leq (\psi - \phi)' \quad \text{on } \overline{\Omega}_+,$$

which contradicts the assumption that $(\psi - \phi)'$ has a strictly negative minimum on Ω_+. This proves that

$$0 \leq (\psi - \phi)' \text{ on } \overline{\Omega}_+,$$

and concludes the proof.

Remark 3.4. In the forward-biased case, the current I is positive. Hence if (2.2) holds, the electric field is positive on $\overline{\Omega}_+$ (whereas it is negative in the reverse-biased case). This explains why we cannot conclude directly in the proof of Theorem 3.3 that $0 \leq (\psi - \phi)'$ on $\overline{\Omega}_+$ (as we did in the proof of Theorem 2.3), and why we have to use in addition a maximum principle argument on $\psi - \phi$. This is also the reason why Theorem 3.3 requires N to satisfy (3.1).

We now study the monotonicity of the current.

THEOREM 3.5. *Assume that* (3.1) *holds and let* V *and* $\tilde{V}$ *be such that* $0 < V < \tilde{V}$. *Assume moreover that every symmetric solution* (ψ, n, p, J_n, J_p) *resp.,* $(\phi, \tilde{n}, \tilde{p}, \tilde{J}_n, \tilde{J}_p)$) *of* (1.1)–$(1.6)_V$ *(resp.,* (1.1)–$(1.6)_{\tilde{V}}$*) satisfies* (2.2). *Then the following property holds:*

$$I < \tilde{I}, \tag{3.6}$$

where

$$I = J_n + J_p \ , \ \tilde{I} = \tilde{J}_n + \tilde{J}_p.$$

Proof. We deduce from the hypotheses and from Theorem 3.3 that (1.1)–$(1.6)_V$ (resp., (1.1)–$(1.6)_{\tilde{V}}$) has a unique symmetric solution.

As in the proof of Theorem 2.8, we claim that (3.6) holds. To the contrary, assume that

$$0 \leq I - \tilde{I}. \tag{3.7}$$

We first prove that if we have

$$(\psi - \phi)'(0) \leq 0 \text{ and } (\psi - \phi)'(1) \leq 0, \tag{3.8}$$

then $(\psi - \phi)' \leq 0$ on $\overline{\Omega}_+$.

Hence assume that (3.8) holds; then if $(\psi - \phi)'$ attains a stictly positive maximum at $y_0 \in \Omega_+$, we deduce, as in the proof of Theorem 3.3, that

$$0 < 2\tilde{I}\varepsilon(\psi - \phi)'(y_0) + 4(np - \tilde{n}\tilde{p})(y_0).$$

By using a maximum principle on $\psi - \phi$ on $(0, y_0)$, we prove as in the proof of Theorem 3.3 that we have a contradiction. Thus we get

$$(\psi - \phi)' \leq 0 \text{ on } \overline{\Omega}_+,$$

which contradicts $V < \tilde{V}$.

Hence we have either $0 < (\psi - \phi)'(0)$ or $0 < (\psi - \phi)'(1)$. We conclude as in the Proof of theorem 2.8 that this case also leads to a contradiction.

Therefore (3.7) is satisfied, and this completes the proof.

3.2. Application to strongly forward-biased symmetric abrupt pn- junctions. We prove in this subsection that the hypotheses of Theorem 3.1 are satisfied for strongly forward-biased pn-junctions.

THEOREM 3.6. *Assume that* (3.1) *holds. Then there exists* $V_0 = V_0(\delta, \varepsilon) > 0$ *such that for all* $V \geq V_0$, (1.1)-$(1.6)_V$ *has a unique solution.*

Moreover, the function

$$V \in [V_0, +\infty) \mapsto I \in \mathbb{R}$$

is continuous and strictly increasing.

From Theorems 3.2, 3.3, and 3.5, it is sufficient to find for which values of V every solution of (1.1)–$(1.6)_V$ satisfies (2.2). As has already been mentioned at the beginning of this section, this property does not hold for all $V > 0$.

However, we will see that this property holds asymptotically; that is, for sufficiently large V. Since the proof is very technical, it is divided into some lemmas.

LEMMA 3.7. *For all* V *satisfying*

$$V \geq \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right) + \frac{\varepsilon}{(n_1+p_1)^2}, \tag{3.9}$$

every solution (ψ, n, p, J_n, J_p) *of* (1.1)–$(1.6)_V$ *satisfies*

$$\psi'(-1) < 0 \quad \text{and} \quad \psi'(1) < 0. \tag{3.10}$$

Proof. First, it is easy to check that $V > 0$ implies $I > 0$. Assume from now on that

$$V > \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right), \tag{3.11}$$

and let us introduce as in [6] the functions

$$m(x) = min(n(x), p(x) - 1), \qquad M(x) = max(n(x), p(x) - 1) \quad \forall x \in [-1, 0],$$

$$\bar{m}(x) = min(n(x) - 1, p(x)), \qquad \bar{M}(x) = max(n(x) - 1, p(x)) \quad \forall x \in [0, 1].$$

One can easily check that [6, §3, Prop. 2] still holds, so that m (resp., $\bar{m}$) has no minima in $(-1, 0)$ (resp., $(0, 1)$) and M (resp., $\bar{M}$) has no maxima in $(-1, 0)$ (resp., $(0, 1)$).

Several cases are now possible:

(i) $0 \leq \psi'(-1)$ and $0 \leq \psi'(1)$,
(ii) $0 > \psi'(-1)$ and $0 \leq \psi'(1)$,
(iii) $0 \leq \psi'(-1)$ and $0 > \psi'(1)$,
(iv) $0 > \psi'(-1)$ and $0 > \psi'(1)$.

Let us first assume that

$$0 \leq \psi'(-1). \tag{3.12}$$

From $I > 0$ and (3.12) we deduce that $m = p - 1$ and $M = n$ in a right neighborhood of -1. Therefore m is decreasing and M is increasing on $(-1, 0)$.

We claim that ψ'' does not vanish on $(-1,0]$. Thus, assume that ψ'' vanishes on $(-1,0]$ and let x_0 be the smallest element of the set

$$\{x \in (-1,0], \psi''(x) = 0\}.$$

Then there exists $x^\star \in (-1, x_0)$ such that $n'(x^\star) = p'(x^\star) = 0$. This implies that $n' < 0$ in a left neighborhood of $x^\star$ ($x^\star$ being excepted), but this contradicts the fact that M is increasing on $(-1,0)$.

Therefore, if (3.12) holds, then

$$0 < \psi''(x) \quad \forall x \in (-1,0]. \tag{3.13}$$

In the same way, we prove that

$$0 \leq \psi'(1), \tag{3.14}$$

implies

$$\psi''(x) < 0 \quad \forall x \in [0,1). \tag{3.15}$$

Hence if case (i) holds, we obtain $\psi(-1) < \psi(1)$, which contradicts (3.11). Thus, case (i) is impossible.

Let us now examine case (ii). We set

$$J = J_n - J_p. \tag{3.16}$$

From (1.2), (1.3), we obtain

$$2(np)' = -I(n-p) + J(n+p). \tag{3.17}$$

We claim that $J > 0$. Thus, assume that $J \leq 0$.

Then, we have $2(np)' \leq -I(n-p)$. Integration of this inequality from -1 to 0 and from 0 to 1 leads to $\psi'(1) < \psi'(-1)$, which contradicts assumption (ii).

Hence $J > 0$ holds. Let now x_0 be the largest element of the set

$$\{x \in (-1,0), \psi'(x) = 0\},$$

which is not empty since (3.15) holds. The maximum principle applied, on the interval $[-1, x_0]$, to the equation

$$\varepsilon\psi^{(3)} = (n+p)\psi' + I, \tag{3.18}$$

subject to Dirichlet boundary conditions at -1 and x_0, leads to

$$\psi' < 0 \quad \text{on } [-1, x_0). \tag{3.19}$$

Hence, ψ has a minimum at $x = x_0$. We now apply the maximum principle, on the interval $[-1, x_0]$, to the equation

$$\varepsilon\psi^{(4)} = (n + p + \varepsilon\psi'^2)\psi'' - \psi'^2 + J\psi',$$

subject to Dirichlet boundary conditions at -1 and x_0.

This yields $0 < \psi''$ on $(-1, x_0)$. Clearly, from (3.18) and from the definition of x_0, we obtain

$$0 \leq \psi''(x_0) < \psi''(x) \leq \psi''(0^-) \quad \text{on } (x_0, 0],$$

and

$$I \leq (n-p)' \quad \text{on } [0,1].$$

Integration of this last inequality from 0 to 1 leads to $\varepsilon\psi''(0^-) \leq 2 - I$, so that we have

$$0 < I < 2. \tag{3.20}$$

From $\psi''(-1) = 0$ and $0 \leq \psi''$ on $[-1, x_0]$, we deduce that $0 < (n-p)'(-1)$.

This, together with the right-hand side of (3.20), implies

$$-\frac{2}{n_1 + p_1} < \psi'(-1) < 0. \tag{3.21}$$

On the other hand, integration of the equation $(n+p)' = (n-p)\psi' + J$ on the intervals $[-1,0]$ and $[0,1]$, and elimination of the unknown constants, gives

$$2\psi(0) = 2J + \frac{\varepsilon}{2}(\psi'(1) + \psi'(-1))(\psi'(1) - \psi'(-1)). \tag{3.22}$$

Using (3.20), (3.21) and $J > 0$ in this last equation, we obtain

$$-\frac{\varepsilon}{(n_1 + p_1)^2} \leq \psi(0).$$

Since (3.11) and $0 < \psi'$ on $[0,1)$ hold, we obtain

$$V < \log\left(\frac{1 + \sqrt{1 + 4\delta^4}}{2\delta^2}\right) + \frac{\varepsilon}{(n_1 + p_1)^2}.$$

Hence, if (3.9) holds, then case (ii) is impossible.

Since case (iii) can be treated in a similar way, the proof of Lemma 3.7 is complete.

Assume now that (3.9) is satisfied, so that (3.10) holds. Then we have either

$$\psi'(0) \leq 0 \quad \text{or} \quad \psi'(0) > 0.$$

If $\psi'(0) \leq 0$, we conclude, from the maximum principle, that $\psi' \leq 0$ on $[-1,1]$, and (2.2) follows at once.

Therefore, the only case left is when (3.10) and $0 < \psi'(0)$ are satisfied. Let us examine this final case.

LEMMA 3.8. *Assume that V satisfies* (3.9) *and let (ψ, n, p, J_n, J_p) be a solution of* (1.1)–(1.6)$_V$, *which is such that* $0 < \psi'(0)$ *and*

$$0 \leq J, \tag{3.23}$$

where J is given by (3.16).

Then the following properties are satisfied:

$$0 \leq \psi'' \quad \text{on } [-1,0], \tag{3.24}$$

$$0 \geq \psi'' \quad \text{on } [0,1]. \tag{3.25}$$

Moreover, there exists a constant

$$V^\star = V^\star(\delta,\varepsilon) > \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right) + \frac{\varepsilon}{(n_1+p_1)^2},$$

such that for all $V \geq V^\star$, *we have*

$$0 < n'(1). \tag{3.26}$$

Proof. From (3.23), we deduce (3.24) as in the proof of lemma 3.7. Now, since

$$(n+p)'(-1) = -\psi'(-1) + J,$$

we have $0 < (n+p)'(-1)$. From this and the inequality $0 < (n-p)'(-1)$, we deduce that $0 < n'(-1)$, which, together with (3.24), implies

$$0 < n' \quad \text{on } [-1,0]. \tag{3.27}$$

On the other hand, (3.17) and (3.23) imply $\psi'(-1) \leq \psi'(1)$, which in turn yields $0 < (n-p)'(1)$. This proves that $\bar{m} = n - 1$ and $\bar{M} = p$ in a left neighborhood of 1. One of the following two cases can now occur:

$$0 \leq p'(1) \tag{3.28}$$

or

$$0 > p'(1). \tag{3.29}$$

Assume first that (3.28) holds. Thus, (3.26) is satisfied. Since $\bar{m}$ has no minima in $(0,1)$, $\bar{m}$ is increasing on $(0,1)$.

Now let y_0 be the smallest element of $\{x \in (0,1), \psi'(x) = 0\}$ and assume that the set $Y = \{x \in (y_0,1), \psi''(x) = 0\}$ is not empty. We call y_1 the largest element of Y. Then, ψ'' has a strictly negative minimum on $(y_1,1)$.

Moreover, we have

$$0 < \psi'(1) + J \leq \psi'(x) + J \leq \psi'(y_1) + J \quad \forall x \in [y_1,1].$$

From the maximum principle applied on the interval $[y_1,1]$ to the equation

$$\varepsilon\psi^{(4)} = (n+p+\varepsilon\psi'^2)\psi'' + \psi'(\psi'+J), \tag{3.30}$$

subject to homogeneous Dirichlet boundary conditions, we obtain a contradiction. Hence we have $Y = \emptyset$, and therefore (3.25) is satisfied.

Assume now that (3.29) holds. Since we have $\bar{M}'(1) = p'(1) < 0$ and since $\bar{M}$ has no maxima on $(0,1)$, $\bar{M}$ is decreasing on $[0,1]$.

We claim that the set $Z = \{x \in (y_0,1], \psi''(x) > 0\}$ is empty. Thus assume that this set is not empty and let y_2 be the first point of $(y_0,1]$, starting from $x = 1$, where ψ'' vanishes by changing its sign.

Since $\bar{M} = p$ on $[y_2,1]$, we have $p'(y_2) \leq 0$. Moreover $\bar{M} = n - 1$ on a left neighborhood of y_2, so that we have $n'(y_2) \leq 0$. From this, we deduce

$$\psi'(y_2) + J = (n+p)'(y_2) \leq 0.$$

On the other hand we deduce from the definition of y_0 that $\psi''(y_0) \leq 0$.

Now let y_3 be the first point, starting from $x = y_2$, where ψ'' vanishes by changing its sign. Then ψ'' has a strictly positive maximum on (y_3, y_2). This contradicts the maximum principle, applied on the interval $[y_3, y_2]$, to (3.30).

Hence $Z = \emptyset$, so that (3.25) holds. We deduce from (3.24) and (3.25) that

$$\psi'(1) + \psi'(-1) \leq 2\psi(1),$$

which together with the inequality $\psi'(-1) \leq \psi'(1)$ implies

$$\psi'(-1) < \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right) - V. \tag{3.31}$$

This, together with the inequality $0 < (n-p)'(-1)$, gives

$$\left(V - \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right)\right)(n_1+p_1) < I. \tag{3.32}$$

In addition, since $0 \leq J$ holds, we obtain

$$\left(V - \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right)\right)(n_1+p_1) < 2J_n. \tag{3.33}$$

Since (3.27) holds, we have $n_{-1} \leq n(0)$. Moreover, since (3.29) and (3.25) hold, we have

$$p' < 0 \quad \text{on } [0,1].$$

Therefore, we obtain

$$p_1^2 \leq n(0)p(0). \tag{3.34}$$

But, since $J \geq 0$ holds, we obtain

$$-I(\varepsilon\psi'' + 1) \leq 2(np)' \quad \text{on } [0,1].$$

Integrating this last equation from 0 to 1, and using (3.34) leads to

$$0 < \varepsilon\psi'(0) \leq 2p_1 I^{-1} + \varepsilon\psi'(1) + 1.$$

We use now (3.32) to obtain

$$-2p_1\left(\left(V - \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right)\right)(n_1+p_1)\varepsilon\right)^{-1} - \frac{1}{\varepsilon} < \psi'(1) < 0. \tag{3.35}$$

We now set $x = 1$ in (1.2) and use (3.33) and (3.35) to deduce that there exists a constant

$$V^\star = V^\star(\delta,\varepsilon) > \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right) + \frac{\varepsilon}{(n_1+p_1)^2},$$

such that for all $V \geq V^\star$, (3.26) holds.

This concludes the proof of Lemma 3.8.

Proof of Theorem 3.6. Let V be such that $V \geq V^\star$ and let (ψ, n, p, J_n, J_p) be a solution of (1.1)–(1.6)$_V$, which is such that $0 < \psi'(0)$.

We assume without loss of generality that (3.23) holds. From (3.25)–(3.27) we obtain

$$n_{-1} \leq n \leq n_1. \tag{3.36}$$

If $0 < p'(1)$ holds, we deduce from (3.25) and from the property that $\bar{M}$ has no maxima in $(0,1)$, that

$$\min(p(0), p(1)) \leq p \leq \max(p(0), p(1)) \quad \text{on } [0,1],$$

from which we obtain

$$n_1 - 2 \leq p \leq n_1 + 1 \quad \text{on } [0,1]. \tag{3.37}$$

In the same way we obtain

$$n_1 - 2 \leq p \leq n_1 + 1 \quad \text{on } [-1,0]. \tag{3.38}$$

If now we have $p'(1) \leq 0$, we prove in the same way that (3.37) and (3.38) hold.

From these inequalities we deduce that

$$|\psi''| \leq \frac{3}{\varepsilon} \quad \text{on } [-1,1].$$

Double integration of this last inequality from -1 to 1 gives

$$V < \frac{3}{2\varepsilon} + \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right).$$

Therefore, if $V \geq V_0$, where

$$V_0 = \max\left(V^\star, \frac{3}{2\varepsilon} + \log\left(\frac{1+\sqrt{1+4\delta^4}}{2\delta^2}\right)\right),$$

then we necessarily have $\psi'(0) \leq 0$.

This proves that for all $V \geq V_0$, every solution of (1.1)–(1.6)$_V$ satisfies (2.2). We now apply Theorems 3.1 and 3.5 to deduce uniqueness and monotonicity of the voltage-current curve on $[V_0, +\infty)$. The continuity of this curve is a direct consequence of this uniqueness result and from [11, Thm. 3.5.2].

Remark 3.9. For all $V \geq V_0$ we proved that the problem (1.1)–(1.6)$_V$ has a unique solution (ψ, n, p, J_n, J_p). Since this solution is symmetric, it is easy to check that it satisfies $(n+p)'(1) \leq 0$. Hence we deduce easily that $p'(1) < 0$ holds for all $V \geq V_0$.

Remark 3.10. From Mock's uniqueness theorem for small $|V|$, we deduce that there exists a constant V_1 in $(0, V_0)$ such that (1.1)–(1.6)$_V$ has a unique solution for all V in $[0, V_1] \cup [V_0, +\infty)$.

4. Conclusion and discussion. We have analyzed in this paper the one-dimensional drift-diffusion model. The uniqueness theorems that are given in the literature for more complex models, as well as for the simplified one under consideration in this paper, hold only for small values of the applied bias. The purpose of the present paper is to show that uniqueness theorems that are valid for arbitrary values of the applied bias can be obtained. Such results are obtained by proving and using new qualitative properties, indeed monotonicity properties, of the drift-diffusion equations. This allows to obtain uniqueness theorems for arbitrary values of the applied bias for the case of symmetric p-n and p-i-n junctions.

We have considered the drift-diffusion model under the assumption of a zero generation-recombination term R, constant mobilities and ideal ohmic contacts. We shall discuss now how our results can be extended to more complex models. The main restriction is due to the assumption that R vanishes. When R does not vanish and depends on the unknowns n and p, as is the case for instance for Shockley–Read–Hall or Auger recombination terms, the electron and hole current densities J_n and J_p are no longer constant, since they satisfy in this case

$$J_n' = R, \qquad J_p' = -R\,.$$

Therefore, if R is not identically zero, $J = J_n - J_p$ is no longer constant, so that the symmetry Theorems 2.2 and 3.2 and the uniqueness Theorems 2.3 and 3.3 do not apply. However, the techniques introduced in this paper for obtaining global uniqueness theorems in the case where $R = 0$, lead (with minor changes) to local uniqueness theorems under the same hypotheses. This, together with the implicit function theorem, give local uniqueness theorems for the case of small generation-recombination terms of the form τR, where τ is a sufficiently small strictly positive parameter. Of course the restriction on τ depends on the applied bias V, which limits the practical use of these results. It would be interesting to have an estimate on the restriction on τ in terms of V. The extension of these results to the case of nonsmall generation-recombination terms is not trivial and requires probably other techniques. Moreover, for technical reasons, our results do not hold for the case of space-dependent mobilities.

As mentioned at the beginning of this section, we have considered the case where the device has ideal ohmic contacts. This means that the boundary data for n and p satisfy the following electro-neutrality condition

$$n_x - p_x - N(x) = 0, \qquad x = \pm 1\,, \tag{4.1}$$

and the following thermal equilibrium condition

$$n_{-1}p_{-1} = n_1 p_1. \tag{4.2}$$

We consider in this paper the case of symmetric boundary data. This means that $\psi_{\pm 1}, n_{\pm 1}, p_{\pm 1}$ satisfy

$$\psi_{-1} = -\psi_1, \quad n_{-1} = p_1, \quad p_{-1} = n_1. \tag{4.3}$$

Therefore, in the case of symmetric boundary data, (4.2) necessarily holds. It is easy to check that the general uniqueness Theorems 2.1 and 3.1 (but also Theorems 2.2, 2.3, 3.1, and 3.2) are still valid for arbitrary boundary data, this even if (4.1) does not hold. These general theorems lead to uniqueness theorems for symmetric p-n and

p-i-n junctions, provided that the solutions of the drift-diffusion model satisfy the a priori conditions (2.2) and (2.3). One can show that these conditions are still satisfied for reverse (resp., forward) biased p-n junctions provided that n_1 and p_1 satisfy

$$N(1) \leq \quad (\text{resp., } \geq)\ n_1 - p_1, \qquad 0 \leq n_1 - p_1. \tag{4.4}$$

so that the uniqueness theorems for reverse (resp., forward) p-n junctions are still valid when (4.1) is replaced by (4.4). Therefore, the uniqueness theorems presented here are valid for more general boundary conditions. We conjecture that they are also still valid for sufficiently large $|V|$, even if (4.4) does not hold. Extension of these results to the case of more general devices (nonsymmetric, multiple junctions, etc.), should also be analyzed.

REFERENCES

[1] F. ALABAU, *Analyse asymptotique et simulation numérique des équations des semiconducteurs*, Ph.D. thesis, Department of Mathematics, University Paris 6, 1987.

[2] ———, *Comportement de la courbe caractéristique potentiel appliqué-courant d'une diode en polarisation inverse et directe*, C.R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 881–886.

[3] ———, *A decoupling method for proving uniqueness theorems for electro-diffusion equations*, in Progress in Partial Differential Equations: Elliptic and Parabolic Problems, Pitman Res. Notes Math. Ser., Vol. 266, Longman Scientific, Harlow, 1992, pp. 107–119.

[4] ———, *A method for proving uniqueness theorems for the stationary semiconductor device and electrochemistry equations*, Nonlinear Anal., 18 (1992), pp. 861–872.

[5] ———, *A uniqueness theorem for reverse-biased diodes*, Applicable Anal., 52 (1994), pp. 261–276.

[6] F. BREZZI, A. CAPELO, AND L. GASTALDI, *A singular perturbation analysis of reverse-biased diodes*, SIAM J. Math. Anal., 20 (1989), pp. 372–387.

[7] A. FRIEDMAN, *Elliptic and parabolic systems associated with semiconductor modeling*, in Progress in Partial Differential Equations: Elliptic and Parabolic Problems, Pitman Res. Notes Math. Ser., Vol. 266, Longman Scientific, Harlow, 1992, pp. 17–23.

[8] J. W. JEROME, *Consistency of semiconductor modelling: An existence/stability analysis for the stationary Van Roosbroeck's system*, SIAM J. Appl. Math., 45 (1985), pp. 565–590.

[9] T. KERKHOVEN, *On the effectiveness of Gummel's method*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 48–60.

[10] ———, *On the one-dimensional current driven semiconductor equations*, SIAM J. Appl. Math., 51 (1991), pp. 748–774.

[11] P. MARKOWICH, *The Stationary Semiconductor Device Equations*, Springer, Wien-New York, 1986.

[12] M. MOCK, *An example of nonuniqueness of stationary solutions in semiconductor device models*, Compel, 1 (1982), pp. 165–174.

[13] ———, *Analysis of Mathematical Models of Semiconductor Devices*, Boole Press, Dublin, 1983.

[14] J. NAUMANN AND M. WOLFF, *A uniqueness theorem for weak solutions of the stationary semiconductor equations*, Appl. Math. Optim., 24 (1991), pp. 223–232.

[15] M. PROTTER AND H. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice Hall, Englewood Cliffs, NJ, 1967.

[16] W. V. ROOSBROECK, *Theory of flow of electrons and holes in germanium and other semiconductors*, Bell Syst. Techn. J., 29 (1950), pp. 560–607.

[17] I. RUBINSTEIN, *Multiple steady-state in one-dimensional electro-diffusion with electroneutrality*, SIAM J. Appl. Math., 47 (1987), pp. 1076–1093.

[18] ———, *Electro-diffusion of ions*, SIAM Stud. in Appl. Math., No. 11, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[19] T. SEIDMAN, *Steady-state solutions of diffusion-reaction systems with electrostatic convection*, Nonlinear Anal., 4 (1980), pp. 623–637.

[20] M. WARD, L. REYNA, AND F. ODEH, *Multiple steady state solutions in a multijunction semiconductor device*, SIAM J. Appl. Math., 51 (1991), pp. 90–123.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 738–749, May 1995

012

LOCAL REGULARITY OF THE ONE-DIMENSIONAL MOTION OF A VISCOELASTIC MEDIUM*

JONG UHN KIM†

Abstract. We establish the local regularity of solutions to the Cauchy problem which arises in linear viscoelasticity. Our method involves MacCamy's trick and Hörmander's result on the propagation of singularity.

Key words. local regularity, viscoelastic medium, microlocal regularity, singular support, bicharacteristic strip

AMS subject classifications. 35B65, 35L99, 45K05, 73F99

Introduction. In this paper we shall investigate the local regularity of solutions to the Cauchy problem associated with the one-dimensional motion of a linear viscoelastic medium. The model equation is given by

$$u_{tt} = a_0(x,t)u_{xx} + a_1(x,t)u_x + a_2(x,t)u_t + a_3(x,t)u + \int_0^t \{b_0(x,t,s)u_{xx}(x,s) + b_1(x,t,s)u_x(x,s)\}\,ds \tag{0-1}$$

for $(x,t) \in R \times [0,\infty)$, where $u(x,t)$ denotes the displacement.

The initial conditions are

$$u(x,0) = u_0(x), \qquad u_t(x,0) = u_1(x), \quad \text{in } R. \tag{0-2}$$

Throughout this paper, we assume that

$$a_j(x,t) \in C^\infty(R \times [0,\infty)) \quad \text{for} \quad j = 0,\dots,3, \tag{0-3}$$

$$\text{all the derivatives of } a_j\text{'s are bounded in } R \times [0,\infty), \tag{0-4}$$

$$a_0(x,t) \geq c > 0 \quad \text{for all} \quad (x,t) \in R \times [0,\infty) \quad \text{for some constant} \quad c, \tag{0-5}$$

$$b_j(x,t,s) \in C^\infty(R \times [0,\infty) \times [0,\infty)) \quad \text{for} \quad j = 0,1, \tag{0-6}$$

$$\text{all the derivatives of } b_j\text{'s are bounded in } R \times [0,\infty) \times [0,\infty). \tag{0-7}$$

There are many mathematical works on dynamic viscoelasticity. Most of them are listed as references in [4] and [11]. In particular, singularity of solution was investigated in [1], [3], [5], [8], [9], and [10] among others. The memory term makes the equation nonlocal and the qualitative behavior of solution depends on the regularity of the memory kernel. A smooth memory kernel can cause the emergence of stationary singularities, which is impossible for hyperbolic equations without memory. On the other hand, a singular kernel can have a regularizing effect on solutions. The precise statements of these phenomena can be found in [8] and [11]. The present work is the outgrowth of an effort to extend some of the results in [8] to an equation with

* Received by the editors March 16, 1992; accepted for publication (in revised form) November 3, 1993.

† Department of Mathematics, Virginia Tech, Blacksburg, Virginia 24061.

variable coefficients. We focus on the smooth memory kernel. The main result is given in Theorem 2.1 below. Our method is different from the previous works. Our argument consists of (i) MacCamy's trick, (ii) Hörmander's result on the propagation of singularity, and (iii) classical argument of the energy method.

MacCamy's trick is to introduce a new unknown function so that the integral term of the equivalent equation involves lower-order derivatives. In order to obtain C^∞ regularity, we repeat differentiation in x and t, respectively, and boost the regularity step by step. This is a typical procedure in the classical energy method. In each step, we use the result of Hörmander [6] on the propagation of singularity to obtain microlocal regularity along each bicharacteristic strip.

In §1, we present some preliminaries and notation. We state our main results in §2 and the proofs are given in §§3 and 4.

1. Notation and preliminaries. We write R_t and R_x if R is the domain of the variables t and x, respectively. When Ω is an open subset of R^n, $n \geq 1$, we employ the standard notation $H^s(\Omega)$, $s \in R$, to represent a Sobolev space. For $u \in \mathcal{D}'(\Omega)$, we say that u is H^s at $x \in \Omega$ if there is a neighborhood $\mathcal{O}$ of x such that $u \in H^s_{loc}(\mathcal{O})$. For $(x_0, \xi_0) \in \Omega \times (R^n \setminus \{0\})$, we say that $(x_0, \xi_0) \notin WF(u)$ if there is a function $\zeta(x) \in C_0^\infty(\Omega)$ which does not vanish at x_0 and a conic neighborhood Γ of ξ_0 such that for each $N \geq 1$,

$$|\mathcal{F}(\zeta u)(\xi)| \leq C_N(1 + |\xi|)^{-N} \qquad \text{for all} \quad \xi \in \Gamma, \tag{1-1}$$

where $\mathcal{F}$ denotes the Fourier transform and C_N is a positive constant depending on N. $WF(u)$ is a closed conic subset of $\Omega \times (R^n \setminus \{0\})$ and is called the wave front set of u. For $u \in \mathcal{D}'(\Omega)$, u is said to be microlocally H^s at $(x_0, \xi_0) \in \Omega \times (R^n \setminus \{0\})$ if we can write $u = u_1 + u_2$, where $u_1 \in H^s_{loc}(\Omega)$ and $(x_0, \xi_0) \notin WF(u_2)$.

2. Statement of the main result and reduction to an equivalent problem. Let $u(x,t)$ be a solution of (0-1) and (0-2) where $(u_0, u_1) \in H^\sigma(R) \times H^{\sigma-1}(R)$, $\sigma \in R$. Our main result is the following.

THEOREM 2.1. *Suppose that* $(x^\star, t^\star) \in R \times (0, \infty)$ *is a point such that each bicharacteristic curve passing through it does not intersect* sing supp $u_0 \cup$ sing supp u_1 *at* $t = 0$. *Then there is a function* $\kappa(x,t) \in C_0^\infty(R^2)$ *such that* $\kappa = 1$ *in a neighborhood of* $(x^\star, t^\star)$ *and*

$$\kappa u \in C^\infty\big(R_t\,;\, H^{\sigma+1}(R_x)\big). \tag{2-1}$$

Furthermore, if $x^\star \notin$ sing supp $u_0 \cup$ sing supp u_1, *in addition we have*

$$\kappa u \in C_0^\infty(R^2). \tag{2-2}$$

Here a curve $x = x(t)$ in the xt-plane is called a bicharacteristic curve if it satisfies

$$\frac{dx}{dt} = \sqrt{a_0(x,t)} \tag{2-3}$$

or

$$\frac{dx}{dt} = -\sqrt{a_0(x,t)}. \tag{2-4}$$

For the proof of this theorem, we reduce equation (0-1) to an equivalent equation whose integral term involves lower-order derivatives. For this, we employ MacCamy's

trick. Let us define $v(x,t)$ by

$$v(x,t) = a_0(x,t)u(x,t) + \int_0^t b_0(x,t,s)u(x,s)\,ds \tag{2-5}$$

for $(x,t) \in R \times [0,\infty)$. Then we can solve (2-5) for u by treating x as a parameter to find

$$u(x,t) = \frac{1}{a_0(x,t)}\,v(x,t) + \int_0^t \beta(x,t,s)v(x,s)\,ds, \tag{2-6}$$

where $\beta \in C^\infty(R \times [0,\infty) \times [0,\infty))$ is determined by a_0 and b_0, and all the derivatives of β are bounded on $R \times [0,T] \times [0,T]$ for each $T > 0$. By substitution into (0-1), we obtain

$$\begin{aligned} v_{tt} = {} & a_0(x,t)v_{xx} + \alpha_1(x,t)v_x + \alpha_2(x,t)v_t + \alpha_3(x,t)v \\ & + \int_0^t \{\beta_1(x,t,s)v_x(x,s) + \beta_2(x,t,s)v(x,s)\}\,ds, \end{aligned} \tag{2-7}$$

where $\alpha_j \in C^\infty(R \times [0,\infty))$, $j = 1,2,3$, and $\beta_j \in C^\infty(R \times [0,\infty) \times [0,\infty))$, $j = 1,2$, are determined by the coefficients of (0-1). All the derivatives of α_j's are bounded in $R \times [0,T]$ and all the derivatives of β_j's are bounded in $R \times [0,T] \times [0,T]$ for each $T > 0$. We present some known facts on the solution of (2-7).

THEOREM 2.2 (existence and uniqueness). *Let $(v_0, v_1) \in H^\sigma(R) \times H^{\sigma-1}(R)$, for $\sigma \in R$. Then there is a unique solution $v(x,t)$ of (2-7) in $C\big([0,\infty)_t\,;\, H^\sigma(R_x)\big) \cap C^1\big([0,\infty)_t\,;\, H^{\sigma-1}(R_x)\big)$ which satisfies*

$$v(x,0) = v_0(x), \qquad v_t(x,0) = v_1(x) \qquad \text{in} \quad R. \tag{2-8}$$

THEOREM 2.3. *Let $v(x,t)$ be the above solution and let $A \in C^\infty(R \times [0,\infty) \times [0,\infty))$ be a function whose derivatives of every order are bounded in $R \times [0,T] \times [0,T]$ for each $T > 0$. Then for each $\zeta(t) \in C_0^\infty((0,\infty))$, it holds that*

$$\zeta(t)v(x,t) \in H^\sigma(R^2) \tag{2-9}$$

and

$$\zeta(t)\int_0^t A(x,t,s)v_x(x,s)\,ds \in H^{\sigma-1}(R^2). \tag{2-10}$$

Proof. When $\sigma \le 0$, (2-9) and (2-10) follow directly from the Fourier transform and the inequality

$$(1 + |\tau| + |\xi|)^\mu \le (1 + |\xi|)^\mu \qquad \text{for all } (\xi,\tau) \in R^2 \text{ and } \mu \le 0. \tag{2-11}$$

When σ is a positive integer, we use

$$\partial_t^k v \in C\big([0,\infty)_t\,;\, H^{\sigma-k}(R_x)\big), \tag{2-12}$$

which follows from (2-7). If $\sigma > 0$ is not an integer, we use the interpolation to get (2-9) and (2-10).

THEOREM 2.4 (domain of dependence). *Let $x = x_+(t)$ and $x = x_-(t)$ denote the bicharacteristic curves so that*

$$\frac{d}{dt}x_\pm(t) = \pm\sqrt{a_0(x_\pm,t)}, \tag{2-13}$$

$$x_\pm(t_0) = x_0, \qquad t_0 > 0. \tag{2-14}$$

If the interval $[x_+(0), x_-(0)]$ *is disjoint from* $\operatorname{supp} v_0 \cup \operatorname{supp} v_1$, *then the solution of* (2-7) *and* (2-8) *vanishes in a neighborhood of* (x_0, t_0).

This can be proved by the argument on pp. 440–448 of [2] when v_0 and v_1 are smooth. In this case, we use the density argument for nonsmooth initial data.

Our assertion on the local regularity of solutions of (2-7) is the following.

THEOREM 2.5. *Let* $v(x,t)$ *be a solution of* (2-7) *and* (2-8). *Suppose that* $(x^\star, t^\star)$ *is a point such that each bicharacteristic curve passing through it does not intersect* $\operatorname{supp} v_0 \cup \operatorname{supp} v_1$ *at* $t = 0$. *Then there is a function* $\kappa(x,t) \in C_0^\infty(R^2)$ *such that* $\kappa = 1$ *in a neighborhood of* $(x^\star, t^\star)$ *and*

$$\kappa v \in C^\infty\big(R_t\,;\, H^{\sigma+2}(R_x)\big). \tag{2-15}$$

Furthermore, if $x^\star \notin \operatorname{supp} v_0 \cup \operatorname{supp} v_1$, *in addition we have*

$$\kappa v \in C_0^\infty(R^2). \tag{2-16}$$

The proof of Theorem 2.5 will be given in §3 and Theorem 2.1 will be proved in §4. We end this section by discussing the effect of the memory term. As mentioned in the introduction, the integral terms in (0-1) and (2-7) make the problem nonlocal. In fact, the integral operators associated with these terms are not even pseudolocal, and more delicate analysis is required for the local regularity. The pseudolocal property is one of the main properties of pseudodifferential operators. To highlight this point, let us consider a simple integral operator $\mathcal{T}$ defined by

$$(\mathcal{T}w)(x,t) = \int_0^t a(x,t,s)w(x,s)\,ds, \tag{2-17}$$

where $a(x,t,s)$ is a smooth function. Without specifying any particular function space, we can easily notice that this integral operator $\mathcal{T}$ is not pseudolocal. For this, it is enough to consider

$$w(x,t) = f(x)g(t), \tag{2-18}$$

where $g(t) \in C_0^\infty(R)$, $g(0) \neq 0$, and $g(t) = 0$ for all $t > 1$. Then it is obvious that $(x^\star, t^\star) \notin \operatorname{sing\,supp} w(x,t)$ for any $t^\star > 1$ and $x^\star \in R$. However, for $t^\star > 1$,

$$(\mathcal{T}w)(x^\star, t^\star) = f(x^\star)\int_0^1 a(x^\star, t^\star, s)g(s)\,ds, \tag{2-19}$$

which can surely have singularity at $(x^\star, t^\star)$ in general. Hence, $\mathcal{T}$ is not pseudolocal in general. It is clear from (2-19) that the singularity of $f(x)$ at $x^\star$ can persist for all future occurrences. This phenomenon is not restricted to the particular functions above. The integral terms in (0-1) and (2-7) can transport the singularity of the initial datum (in a milder form) along the straight line parallel to the t axis in the xt-plane. The analytical mechanism for this is explicitly exposed in (3-34) and (4-5) below, where the initial data have gotten out of the integrals to influence the local regularity in the future.

3. Proof of Theorem 2.5. Let $v(x,t)$ satisfy (2-7) in $R \times (0,\infty)$. If we differ-

entiate (2-7) in t, we obtain

$$
\begin{aligned}
(3\text{-}1)\qquad (\partial_t v)_{tt} &= a_0(\partial_t v)_{xx} + \alpha_1(\partial_t v)_x + \left(\alpha_2 + \frac{1}{a_0}\,\partial_t a_0\right)(\partial_t v)_t \\
&+ \left(\partial_t\alpha_1 - \frac{\alpha_1}{a_0}\,\partial_t a_0 + \beta_1(x,t,t)\right) v_x \\
&+ \left(\partial_t\alpha_2 - \frac{\alpha_2}{a_0}\,\partial_t a_0 + \alpha_3\right) v_t \\
&+ \left(\partial_t\alpha_3 + \beta_2(x,t,t) - \frac{\alpha_3}{a_0}\,\partial_t a_0\right) v \\
&+ \int_0^t \Big\{ \partial_t\beta_1(x,t,s)v_x(x,s) + \partial_t\beta_2(x,t,s)v(x,s) \\
&\qquad - \frac{1}{a_0(x,t)}\,\beta_1(x,t,s)\partial_t a_0(x,t)v_x(x,s) \\
&\qquad - \frac{1}{a_0(x,t)}\,\beta_2(x,t,s)\partial_t a_0(x,t)v(x,s)\Big\}\,ds,
\end{aligned}
$$

where we have expressed v_{xx} in terms of other derivatives and an integral by using (2-7). By induction, we can derive for $k \geq 1$

$$
\begin{aligned}
(3\text{-}2)\qquad (\partial_t^k v)_{tt} &= a_0(\partial_t^k v)_{xx} + \alpha_1(\partial_t^k v)_x \\
&+ (\alpha_2 + \frac{k}{a_0}\,\partial_t a_0)(\partial_t^k v)_t \\
&+ b_{k,1}(\partial_t^{k-1} v)_x + \cdots + b_{k,k}\,v_x \\
&+ c_{k,0}\,\partial_t^k v + c_{k,1}\,\partial_t^{k-1} v + \cdots + c_{k,k}\,v \\
&+ \int_0^t \{\gamma_{k,1}(x,t,s)v_x(x,s) + \gamma_{k,2}(x,t,s)v(x,s)\}\,ds,
\end{aligned}
$$

where a_0, α_1 and α_2 are the same as in (2-7) and the remaining coefficients depend on k. It is easy to see that $b_{k,j}$'s, $c_{k,j}$'s, and all of their derivatives are bounded in $R \times [0, T]$ for each $T > 0$, and that $\gamma_{k,1}$, $\gamma_{k,2}$, and all of their derivatives are bounded in $R \times [0, T] \times [0, T]$ for each $T > 0$. Equation (3-2) will be the basic identity for the proof of Theorem 2.5.

Next we set

$$(3\text{-}3)\qquad \sigma(x,t) = \sqrt{a_0(x,t)},$$

$$(3\text{-}4)\qquad p_1(x,t,\xi,\tau) = \tau + \sigma(x,t)\xi,$$

$$(3\text{-}5)\qquad p_2(x,t,\xi,\tau) = \tau - \sigma(x,t)\xi.$$

For $j = 1, 2$, a bicharacteristic strip of p_j is defined to be a solution $(x(t), t, \xi(t), \tau(t))$ of

$$(3\text{-}6)\qquad \frac{dx}{dt} = \frac{\partial p_j}{\partial \xi},$$

$$(3\text{-}7)\qquad \frac{d\xi}{dt} = -\frac{\partial p_j}{\partial x},$$

$$\frac{d\tau}{dt} = -\frac{\partial p_j}{\partial t}, \tag{3-8}$$

which satisfies

$$p_j(x(t), t, \xi(t), \tau(t)) = 0, \tag{3-9}$$

$$(\xi(t),\ \tau(t)) \neq (0,\ 0). \tag{3-10}$$

Since $\sigma(x,t)$ and its derivatives are bounded in $R \times [0,\infty)$, each bicharacteristic strip exists globally for $t \geq 0$. The curve in the xt-plane described by the above $x(t)$ is called a bicharacteristic curve, which was already defined by (2-3) or (2-4).

Next we choose any $(x^\star, t^\star)$ such that $t^\star > 0$ and each bicharacteristic curve passing through $(x^\star, t^\star)$ does not intersect supp $v_0 \cup$ supp v_1 at $t = 0$. For $j = 1, 2$, let $\Gamma_j(t) = (x_j(t), t, \xi_j(t), \tau_j(t))$ stand for a bicharacteristic strip of $p_j(x,t,\xi,\tau)$ such that

$$x_j(t^\star) = x^\star. \tag{3-11}$$

Then, by virtue of Theorem 2.4, there is a positive constant $\epsilon^\star$ such that

$$\text{the set } \{(x_j(t),\ t) : 0 \leq t \leq \epsilon^\star\} \bigcap \text{supp } v(x,t) \quad \text{is empty} \quad \text{for } j = 1, 2, \tag{3-12}$$

where $v(x,t)$ is a solution of (2-7) and (2-8). We will first obtain microlocal regularity in a conic neighborhood of $\Gamma_j(t)$, $\epsilon^\star \leq t \leq t^\star$ for $j = 1, 2$.

LEMMA 3.1. *For each $k \geq 0$, $j = 1, 2$, it holds that*

$$\partial_t^k v \text{ is microlocally } H^\sigma \text{ on } \Gamma_j(t),\ \epsilon^\star \leq t \leq t^\star. \tag{3-13}$$

Proof. (3-13) is true for $k = 0$ by Theorem 2.3. Suppose that (3-13) is true for $k = 0, 1, \ldots, m-1$, and set

$$\Phi = \partial_t^m v. \tag{3-14}$$

Then it follows from (3-2) that

$$\begin{aligned} \Phi_{tt} = a_0 \Phi_{xx} + \alpha_1 \Phi_x + \left(\alpha_2 + \frac{m}{a_0}\,\partial_t a_0\right) \Phi_t \\ + f_{1,m} + f_{2,m}, \end{aligned} \tag{3-15}$$

where

$$\begin{aligned} f_{1,m} = b_{m,1}(\partial_t^{m-1} v)_x + \cdots + b_{m,m}\, v_x \\ + c_{m,0}\, \partial_t^m v + c_{m,1}\, \partial_t^{m-1} v + \cdots + c_{m,m}\, v, \end{aligned} \tag{3-16}$$

$$f_{2,m} = \int_0^t \{\gamma_{m,1}(x,t,s) v_x(x,s) + \gamma_{m,2}(x,t,s) v(x,s)\}\, ds. \tag{3-17}$$

Since (3-13) is valid for $k = 0, 1, \ldots, m-1$, we use Theorem 18.1.31 of [7] to find that

$$f_{1,m} \text{ is microlocally } H^{\sigma-1} \text{ on } \Gamma_j(t) \text{ for } \epsilon^\star \leq t \leq t^\star,\ j = 1, 2. \tag{3-18}$$

Next it follows from Theorem 2.3 that

$$f_{2,m} \text{ is microlocally } H^{\sigma-1} \text{ on } \Gamma_j(t) \text{ for } \epsilon^\star \leq t \leq t^\star,\ j = 1, 2. \tag{3-19}$$

By virtue of (3-12), it is apparent that

$$\Phi \text{ is microlocally } H^\sigma \text{ at } \Gamma_j(\epsilon^\star),\ j=1,2. \tag{3-20}$$

Now Proposition 3.5.1 of [6] yields that

$$\Phi \text{ is microlocally } H^\sigma \text{ on } \Gamma_j(t) \text{ for } \epsilon^\star \le t \le t^\star,\ j=1,2. \tag{3-21}$$

By induction, this completes the proof.

Next we obtain local regularity.

LEMMA 3.2. *For each $k \ge 0$, it holds that*

$$\partial_t^k v \quad \text{is} \quad H^\sigma \quad \text{at} \quad (x^\star, t^\star). \tag{3-22}$$

Proof. For $k=0$, (3-22) is true by Theorem 2.3. Suppose that (3-22) is valid for $k=0,1,\dots,m-1$. We define Φ by (3-14) and rewrite (3-15) as

$$\left(\partial_{tt} - a_0\partial_{xx} - \alpha_1\partial_x - \left(\alpha_2 + \frac{m}{a_0}\,\partial_t a_0\right)\partial_t\right)\Phi = f_{1,m} + f_{2,m}. \tag{3-23}$$

Since (3-22) is true for $k=0,1,\dots,m-1$, it holds that

$$f_{1,m} \quad \text{is} \quad H^{\sigma-1} \quad \text{at} \quad (x^\star, t^\star). \tag{3-24}$$

Again by Theorem 2.3, we find that

$$f_{2,m} \quad \text{is} \quad H^{\sigma-1} \quad \text{at} \quad (x^\star, t^\star). \tag{3-25}$$

Now let us choose any $(\xi_0, \tau_0) \ne (0,0)$ such that

$$\tau_0^2 \ne a_0(x^\star, t^\star)\,\xi_0^2. \tag{3-26}$$

It follows from Theorem 18.1.31 of [7] that

$$\Phi \quad \text{is microlocally} \quad H^{\sigma+1} \quad \text{at} \quad (x^\star, t^\star, \xi_0, \tau_0). \tag{3-27}$$

Combining (3-13) and (3-27), we find that for every $(\xi,\tau) \ne (0,0)$,

$$\Phi \quad \text{is microlocally} \quad H^\sigma \quad \text{at} \quad (x^\star, t^\star, \xi, \tau), \tag{3-28}$$

which yields that

$$\Phi \quad \text{is} \quad H^\sigma \quad \text{at} \quad (x^\star, t^\star). \tag{3-29}$$

By induction, the proof is complete.

For the above $(x^\star, t^\star)$, we can find a positive number $r < t^\star$ such that each bicharacteristic curve that meets the ball $B_r((x^\star, t^\star))$ does not intersect supp $v_0 \cup$ supp v_1 at $t=0$.

Then we can apply Lemma 3.2 to each $(x,t) \in B_r((x^\star, t^\star))$ so that

$$\partial_t^k v \in H^\sigma_{\mathrm{loc}}\big(B_r((x^\star, t^\star))\big) \qquad \text{for each} \quad k \ge 0. \tag{3-30}$$

We note that r depends on $(x^\star, t^\star)$, but is independent of k.

Next we will raise the local regularity. For this we need the following fact.

LEMMA 3.3. *Let $f_{2,m}$ be defined by* (3-17). *Then, for each $m \ge 1$,*

$$f_{2,m} \in C\big([0,\infty)_t\,;\, H^\sigma(R_x)\big). \tag{3-31}$$

Proof. We already know that

$$f_{2,m} \in C\big([0,\infty)_t\,;\, H^{\sigma-1}(R_x)\big). \tag{3-32}$$

Let us consider $\partial_x f_{2,m}$:

(3-33)
$$\partial_x f_{2,m}(x,t) = \int_0^t \{(\partial_x \gamma_{m,1}(x,t,s))v_x(x,s) + (\partial_x \gamma_{m,2}(x,t,s))v(x,s) + \gamma_{m,1}(x,t,s)v_{xx}(x,s) + \gamma_{m,2}(x,t,s)v_x(x,s)\}\, ds.$$

Since v is a solution of (2-7), we can write

(3-34)
$$\begin{aligned}
&\int_0^t \gamma_{m,1}(x,t,s)v_{xx}(x,s)\, ds \\
&= \int_0^t \big(\gamma_{m,1}(x,t,s)/a_0(x,s)\big)\big[v_{ss}(x,s) - \alpha_1 v_x(x,s) - \alpha_2 v_s(x,s) - \alpha_3 v(x,s) \\
&\qquad - \int_0^s \{\beta_1(x,s,\eta)v_x(x,\eta) + \beta_2(x,s,\eta)v(x,\eta)\}\, d\eta\big]\, ds \\
&= \big(\gamma_{m,1}(x,t,s)/a_0(x,s)\big)\{v_s(x,s) - \alpha_2(x,s)v(x,s)\}\Big|_{s=0}^{s=t} \\
&- \partial_s\big(\gamma_{m,1}(x,t,s)/a_0(x,s)\big)v(x,s)\Big|_{s=0}^{s=t} \\
&+ \int_0^t \{\partial_{ss}\big(\gamma_{m,1}(x,t,s)/a_0(x,s)\big) + \partial_s\big(\gamma_{m,1}(x,t,s)\alpha_2(x,s)/a_0(x,s)\big)\}v(x,s)\, ds \\
&+ \mathcal{R}(x,t).
\end{aligned}$$

Here, $\mathcal{R}(x,t)$ is given by

(3-35)
$$\begin{aligned}
\mathcal{R}(x,t) = &- \int_0^t \big(\gamma_{m,1}(x,t,s)/a_0(x,s)\big)\big[\alpha_1 v_x(x,s) + \alpha_3 v(x,s)\big]\, ds \\
&- \int_0^t \{\rho_1(x,t,s)v_x(x,s) + \rho_2(x,t,s)v(x,s)\}\, ds,
\end{aligned}$$

where

(3-36)
$$\rho_1(x,t,s) = \int_s^t \{\gamma_{m,1}(x,t,\eta)\beta_1(x,\eta,s)/a_0(x,\eta)\}\, d\eta,$$

(3-37)
$$\rho_2(x,t,s) = \int_s^t \{\gamma_{m,1}(x,t,\eta)\beta_2(x,\eta,s)/a_0(x,\eta)\}\, d\eta.$$

It is now evident that

(3-38)
$$\int_0^t \gamma_{m,1}(x,t,s)v_{xx}(x,s)\, ds \in C\big([0,\infty)_t\,;\, H^{\sigma-1}(R_x)\big)$$

and consequently,

(3-39)
$$\partial_x f_{2,m}(x,t) \in C\big([0,\infty)_t\,;\, H^{\sigma-1}(R_x)\big),$$

which yields (3-31).

Now (3-30) implies that

$$v \in C^\infty\left(\left[t^\star - \frac{r}{2}, t^\star + \frac{r}{2}\right]; H^\sigma\left(\left(x^\star - \frac{r}{2}, x^\star + \frac{r}{2}\right)\right)\right), \tag{3-40}$$

from which it follows that

$$f_{1,m} \in C^\infty\left(\left[t^\star - \frac{r}{2}, t^\star + \frac{r}{2}\right]; H^{\sigma-1}\left(\left(x^\star - \frac{r}{2}, x^\star + \frac{r}{2}\right)\right)\right). \tag{3-41}$$

Let us rewrite (3-15) as

$$(a_0\partial_{xx} + \alpha_1\partial_x)\Phi = \Phi_{tt} - \left(\alpha_2 + \frac{m}{a_0}\,\partial_t a_0\right)\Phi_t - f_{1,m} - f_{2,m}. \tag{3-42}$$

By (3-31) and (3-41), we can infer from (3-42)

$$\Phi \in C\left(\left[t^\star - \frac{r}{2}, t^\star + \frac{r}{2}\right]; H^{\sigma+1}_{loc}\left(\left(x^\star - \frac{r}{2}, x^\star + \frac{r}{2}\right)\right)\right). \tag{3-43}$$

Since m is arbitrary, we have

$$v \in C^\infty\left(\left[t^\star - \frac{r}{2}, t^\star + \frac{r}{2}\right]; H^{\sigma+1}_{loc}\left(\left(x^\star - \frac{r}{2}, x^\star + \frac{r}{2}\right)\right)\right), \tag{3-44}$$

which, in turn, yields

$$f_{1,m} \in C^\infty\left(\left[t^\star - \frac{r}{2}, t^\star + \frac{r}{2}\right]; H^{\sigma}_{loc}\left(\left(x^\star - \frac{r}{2}, x^\star + \frac{r}{2}\right)\right)\right). \tag{3-45}$$

Again by (3-31), (3-42), and (3-45), we derive (2-15).

Next we further assume

$$x^\star \notin \text{supp } v_0 \cup \text{supp } v_1. \tag{3-46}$$

Then there is a positive number $\rho < \frac{1}{2}r$ such that

$$\text{distance } (x^\star,\ \text{supp } v_0 \cup \text{supp } v_1) > \rho. \tag{3-47}$$

We need the following identity for $f_{2,m}$ which was defined by (3-17).

LEMMA 3.4. *For $|x - x^\star| < \rho$ and $t > 0$, it holds that for each $N \geq 1$,*

$$\begin{aligned} \partial_x^N f_{2,m}(x,t) &= d_{m,N,N-1}\,\partial_x^{N-1} v_t(x,t) + \cdots + d_{m,N,0}\, v_t(x,t) \\ &\quad + e_{m,N,N-1}\,\partial_x^{N-1} v(x,t) + \cdots + e_{m,N,0}\, v(x,t) \\ &\quad + \int_0^t \{g_{m,N,1}(x,t,s)v_x(x,s) + g_{m,N,2}(x,t,s)v(x,s)\}\,ds, \end{aligned} \tag{3-48}$$

where the coefficients depend on m and N and are infinitely differentiable.

Proof. For $N = 1$, we recall (3-33) and (3-34). By virtue of (3-47), we have

$$v(x,0) = v_t(x,0) = 0 \qquad \text{for} \quad |x - x^\star| < \rho. \tag{3-49}$$

Hence, (3-48) is valid for $N = 1$. For $N \geq 2$, we repeat the same argument as for (3-34) to establish (3-48) by induction.

We now differentiate (3-15) in x N times to obtain

(3-50)
$$a_0\,\partial_x^{N+2}\,\partial_t^m\,v(x,t) = \sum_{k=0}^{m+2}\sum_{j=0}^{N+1} h_{j,k}^{m,N}\,\partial_x^j\,\partial_t^k\,v(x,t) - \int_0^t \{g_{m,N,1}(x,t,s)v_x(x,s) + g_{m,N,2}(x,t,s)v(x,s)\}\,ds,$$

for $|x-x^\star| < \rho$, $t > 0$, where the coefficients $h_{j,k}^{m,N}$'s are infinitely differentiable. Since we have

(3-51)
$$\partial_t^m\,v \in C\Big(\Big[t^\star - \frac{r}{2},\, t^\star + \frac{r}{2}\Big];\, H_{loc}^{\sigma+2}\Big(\Big(x^\star - \frac{r}{2},\, x^\star + \frac{r}{2}\Big)\Big)\Big),$$

for every $m \geq 0$, it follows that

(3-52)
$$\sum_{k=0}^{m+2}\sum_{j=0}^{N+1} h_{j,k}^{m,N}\,\partial_x^j\,\partial_t^k\,v \in C\Big(\Big[t^\star - \frac{r}{2},\, t^\star + \frac{r}{2}\Big];\, H_{loc}^{\sigma+1-N}\Big(\Big(x^\star - \frac{r}{2},\, x^\star + \frac{r}{2}\Big)\Big)\Big).$$

By the same argument as for (3-31), we can easily see that

(3-53)
$$\int_0^t \{g_{m,N,1}(x,t,s)v_x(x,s) + g_{m,N,2}(x,t,s)v(x,s)\}\,ds \in C\big([0,\infty)_t;\, H^\sigma(R_x)\big).$$

By virtue of (3-50), (3-52), and (3-53), we find that

(3-54)
$$\partial_t^m\,v \in C\Big(\Big[t^\star - \frac{r}{2},\, t^\star + \frac{r}{2}\Big];\, H_{loc}^{\sigma+3}\Big(\Big(x^\star - \rho,\, x^\star + \rho\Big)\Big)\Big),$$

for every $m \geq 0$, which, in turn, yields

(3-55)
$$\sum_{k=0}^{m+2}\sum_{j=0}^{N+1} h_{j,k}^{m,N}\,\partial_x^j\,\partial_t^k\,v \in C\Big(\Big[t^\star - \frac{r}{2},\, t^\star + \frac{r}{2}\Big];\, H_{loc}^{\sigma+2-N}\Big(\Big(x^\star - \rho,\, x^\star + \rho\Big)\Big)\Big).$$

If $N \geq 2$, then (3-50) implies that

(3-56)
$$\partial_t^m\,v \in C\Big(\Big[t^\star - \frac{r}{2},\, t^\star + \frac{r}{2}\Big];\, H_{loc}^{\sigma+4}\Big(\Big(x^\star - \rho,\, x^\star + \rho\Big)\Big)\Big),$$

for every $m \geq 1$. We can continue this process until we arrive at

(3-57)
$$\partial_t^m\,v \in C\Big(\Big[t^\star - \frac{r}{2},\, t^\star + \frac{r}{2}\Big];\, H_{loc}^{\sigma+N+2}\Big(\Big(x^\star - \rho,\, x^\star + \rho\Big)\Big)\Big).$$

Since N is arbitrary, we have proved (2-16).

4. Proof of Theorem 2.1. According to (2-5), we have

(4-1)
$$v_0(x) = a_0(x,0)u_0(x),$$

(4-2)
$$v_1(x) = a_0(x,0)u_1(x) + \partial_t a_0(x,0)u_0(x) + b(x,0,0)u_0(x).$$

Suppose that $(x^\star, t^\star)$ is a point such that $t^\star > 0$ and each bicharacteristic curve passing through it does not meet supp $u_0 \cup$ supp u_1 at $t = 0$. Hence, each bicharacteristic curve passing through $(x^\star, t^\star)$ does not meet supp $v_0 \cup$ supp v_1 at $t = 0$, so

we can apply Theorem 2.5. Then, by means of (2-6), we obtain the local regularity of $u(x,t)$. For this, we consider the local regularity of the integral in (2-6).

LEMMA 4.1. *For the above* $(x^\star, t^\star)$, *let* $r > 0$ *be the same as in* (3-44). *Then, it holds that for each* $m \geq 0$,

$$\partial_t^m \int_0^t \beta(x,t,s)v(x,s)\,ds \in C\Big(\Big[t^\star - \frac{r}{2},\, t^\star + \frac{r}{2}\Big];\, H^{\sigma+1}_{loc}\Big(\Big(x^\star - \frac{r}{2},\, x^\star + \frac{r}{2}\Big)\Big)\Big). \tag{4-3}$$

Proof. By induction, it is easy to see that

$$\begin{aligned}\partial_t^m \int_0^t \beta(x,t,s)v(x,s)\,ds &= q_{m,1}(x,t)\partial_t^{m-1}v(x,t) + \cdots + q_{m,m}(x,t)v(x,t)\\ &\quad + \int_0^t J_m(x,t,s)v(x,s)\,ds,\end{aligned} \tag{4-4}$$

where all the derivatives of $q_{m,j}$'s are bounded in $R \times [0,\, T]$ and all the derivatives of J_m are bounded in $R \times [0,\, T] \times [0,\, T]$ for each $T > 0$. Through the same procedure as for (3-34), we find that

$$\begin{aligned}\partial_{xx} \int_0^t J_m(x,t,s)v(x,s)\,ds &= \frac{1}{a_0(x,t)} J_m(x,t,t)v_t(x,t)\\ &\quad - \frac{1}{a_0(x,0)} J_m(x,t,0)v_1(x) + \mathcal{R}_{m,1} + \mathcal{R}_{m,2}.\end{aligned} \tag{4-5}$$

Here $\mathcal{R}_{m,1}$ is a linear combination of $v(x,t)$ and $v_0(x)$ with smooth coefficient functions so that $\mathcal{R}_{m,1}$ belongs to $C\big([0,\infty)_t;\, H^\sigma(R_x)\big)$, and $\mathcal{R}_{m,2}$ consists of the integrals similar to (3-35), which obviously belong to $C\big([0,\infty)_t;\, H^{\sigma-1}(R_x)\big)$. Then, by differentiating the integrals in x and integrating by parts as in (3-33) and (3-34), we can easily show that $\mathcal{R}_{m,2}$ belongs to $C\big([0,\infty)_t;\, H^\sigma(R_x)\big)$. It now follows from (4-5) that

$$\int_0^t J_m(x,t,s)v(x,s)\,ds \in C\big([0,\infty)_t;\, H^{\sigma+1}(R_x)\big). \tag{4-6}$$

Hence, we obtain (4-3) from (3-44), (4-4), and (4-6).

Now (2-1) follows from (2-6), (3-44), and (4-3).

Remark 4.2. If $u_0 \in H^\sigma(R)$ and $u_1 \in H^\sigma(R)$, then it is easy to replace (2-1) by

$$\kappa u \in C^\infty\big(R_t;\, H^{\sigma+2}(R_x)\big). \tag{4-7}$$

This is due to the improved regularity

$$v_1 \in H^\sigma(R), \tag{4-8}$$

which together with (4-5) yields

$$\int_0^t J_m(x,t,s)v(x,s)\,ds \in C\Big(\Big[t^\star - \frac{r}{2},\, t^\star + \frac{r}{2}\Big];\, H^{\sigma+2}_{loc}\Big(\Big(x^\star - \frac{r}{2},\, x^\star + \frac{r}{2}\Big)\Big)\Big). \tag{4-9}$$

Then, by (2-15) and (4-9), we get (4-7).

It remains to prove (2-2). If $x \notin \text{supp } u_0 \cup \text{supp } u_1$, we have the following identity. For all $m \geq 0$ and $N \geq 0$,

(4-10)

$$\partial_x^N \partial_t^m u(x,t) = \sum_{j=0}^{m+1} \sum_{k=0}^{N} B_{j,k}^{m,N}(x,t)\, \partial_t^j \partial_x^k v(x,t) + \int_0^t \{\psi_{m,N,1}(x,t,s) v_x(x,s) + \psi_{m,N,2}(x,t,s) v(x,s)\}\, ds,$$

where all the derivatives of $B_{j,k}^{m,N}$'s are bounded in $R \times [0, T]$, and all the derivatives of $\psi_{m,N,1}$ and $\psi_{m,N,2}$ are bounded in $R \times [0, T] \times [0, T]$ for each $T > 0$. This identity can be proved by induction. We can easily derive (2-2) from (2-16) and (4-10). Since our problem is linear, we can use superposition and the well-known regularity of solutions with smooth initial data to replace the condition $x^\star \notin \text{supp } u_0 \cup \text{supp } u_1$ by the condition $x^\star \notin \text{sing supp } u_0 \cup \text{sing supp } u_1$. Now the proof of Theorem 2.1 is complete.

Acknowledgment. The author thanks Professors K. Hannsgen, M. Renardy, and R. Wheeler for useful information on this subject.

REFERENCES

[1] B. D. Coleman, M. E. Gurtin, and I. R. Herrera, *Waves in materials with memory*, Arch. Rational Mech. Anal., 19 (1965), pp. 1–19; 239–265.

[2] R. Courant and D. Hilbert, *Methods of Mathematical Physics* Vol. 2, Interscience Publishers, New York, London, Sydney, 1966.

[3] J. M. Greenberg, L. Hsiao, and R. C. MacCamy, *A model Riemann problem for Volterra equations*, in Volterra and Functional Differential Equations, K. Hannsgen et al., eds., Marcel Dekker, New York, 1982, pp. 25–43.

[4] G. Gripenberg, S-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, New York, Sydney, 1990.

[5] K. B. Hannsgen and R. L. Wheeler, *Behavior of the solutions of a Volterra equation as a parameter tends to infinity*, J. Integral Equations, 7 (1984), pp. 229–237.

[6] L. Hörmander, *On the existence and the regularity of solutions of linear pseudo-differential equations*, L'Enseignement Math., 17 (1971), pp. 99–163.

[7] ———, *The Analysis of Linear Partial Differential Operators*, Vol. 3, Springer-Verlag, Berlin, Heidelberg, New York, 1985.

[8] W. J. Hrusa and M. Renardy, *On wave propagation in linear viscoelasticity*, Quart. Appl. Math., 43 (1985), pp. 237–254.

[9] R. C. MacCamy, *A model Riemann problem for Volterra equations*, Arch. Rational Mech. Anal., 82 (1983), pp. 71–86.

[10] M. Renardy, *Some remarks on the propagation and non-propagation of discontinuities in linearly viscoelastic liquids*, Rheol. Acta, 21 (1982), pp. 251–254.

[11] M. Renardy, W. J. Hrusa, and J. A. Nohel, *Mathematical Problems in Viscoelasticity*, Longman, New York, 1987.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 750–760, May 1995

013

OSCILLATIONS OF SOLUTIONS TO THE TWO-DIMENSIONAL BROADWELL MODEL, AN H-MEASURE APPROACH*

ROBERT PESZEK†

Abstract. We study oscillatory sequences of solutions $u_1^\epsilon, u_2^\epsilon, u_3^\epsilon, u_4^\epsilon$ to the two-dimensional Broadwell model. We ask the following question: What will happen if at the initial time $u_2^\epsilon, u_3^\epsilon$, and u_4^ϵ converge strongly while u_1^n is left to oscillate? Can oscillations be created in $u_2^\epsilon, u_3^\epsilon$, or u_4^ϵ at later times? It turns out that the answer depends on the direction in which the oscillations occur. We apply the H-measures to study this problem.

Key words. Broadwell model, H-measure, oscillations, Young measure

AMS subject classifications. 35L45, 35B05, 76P05

1. Introduction.

In recent years, there has been a considerable interest in studying oscillatory sequences (i.e., sequences that converge weakly but do not converge strongly in L^1_{loc}) of solutions to nonlinear hyperbolic systems. Most of the results obtained were proven with the aid of compensated compactness theory and Young measures. Unfortunately, the successful applications of these theories are restricted to one-dimensional cases. Recent studies ([1], [8], [9]) tend to confirm the fact that one needs new tools, other than Young measures, to attack multidimensional hyperbolic systems. Some of such tools already have been created (semiclassical measures of Gerard [8], Wigner measures of Lions and Paul [9], and H-measures introduced by Tartar [1]).

This note deals with propagation of oscillations in the two-dimensional (2D) Broadwell model with the aid of H-measures. The results presented here can be extended to the three-dimensional (3D) Broadwell model and to other similar models. For simplicity we restrict our attention to the 2D case and consider the following system of partial differential equations:

$$
\begin{aligned}
&\frac{\partial}{\partial t}u_1(x,y,t) + \frac{\partial}{\partial x}u_1(x,y,t) = u_3(x,y,t)u_4(x,y,t) - u_1(x,y,t)u_2(x,y,t),\\
&\frac{\partial}{\partial t}u_2(x,y,t) - \frac{\partial}{\partial x}u_2(x,y,t) = u_3(x,y,t)u_4(x,y,t) - u_1(x,y,t)u_2(x,y,t),\\
&\frac{\partial}{\partial t}u_3(x,y,t) + \frac{\partial}{\partial y}u_3(x,y,t) = u_1(x,y,t)u_2(x,y,t) - u_3(x,y,t)u_4(x,y,t),\\
&\frac{\partial}{\partial t}u_4(x,y,t) - \frac{\partial}{\partial y}u_4(x,y,t) = u_1(x,y,t)u_2(x,y,t) - u_3(x,y,t)u_4(x,y,t),
\end{aligned}
\tag{1}
$$

with standard initial conditions imposed at time $t = 0$.

System (1) and other similar systems were studied by many authors. We refer the reader to the review paper of Płatkowski and Illner [7] and to the references contained therein for a survey of the theory of such systems. Equations (1) model a motion of an idealized gas of particles that can travel only with prescribed velocities. Specifically, u_1, u_2, u_3, and u_4 represent the number (or the density) of particles that travel with velocities $(1,0)$, $-(1,0)$, $(0,1)$, and $-(0,1)$; $\pm(u_3u_4 - u_1u_2)$ are the collision terms; and equations (1) are simply the balance identities for such a gas.

The compensated compactness theory was applied to such models in the one-dimensional (1D) context by Tartar ([2], [3]) who has studied interaction of oscillations

* Received by the editors May 10, 1993; accepted for publication October 25, 1993. This research was conducted during the author's postdoctoral appointment at the Department of Mathematics and the Center for Nonlinear Analysis, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, and was partially supported by the U.S. Army Research Office.

† Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan 49931.

in 1D semilinear hyperbolic systems. In particular, Tartar has considered the following reduction of the system (1):

$$\begin{aligned} &\tfrac{\partial}{\partial t}u_1(x,t) + \tfrac{\partial}{\partial x}u_1(x,t) = u_3^2(x,t) - u_1(x,t)u_2(x,t),\\ &\tfrac{\partial}{\partial t}u_2(x,t) - \tfrac{\partial}{\partial x}u_2(x,t) = u_3^2(x,t) - u_1(x,t)u_2(x,t),\\ &\tfrac{\partial}{\partial t}u_3(x,t) = u_1(x,t)u_2(x,t) - u_3^2(x,t) \end{aligned} \tag{2}$$

obtained by assuming that u_1, u_2, u_3, and u_4 are independent of y and $u_3 = u_4$. Assume that $(u_1^\epsilon, u_2^\epsilon, u_3^\epsilon)$ is a sequence of solutions to (2) and that $(u_1^\epsilon, u_2^\epsilon, u_3^\epsilon)$ lies in a bounded set of L^∞. Let σ_1^2, σ_2^2, and σ_3^2 denote the variances of the Young measures generated by the sequences $u_1^\epsilon, u_2^\epsilon$, and u_3^ϵ, respectively. Tartar analysis shows that

$$\begin{aligned} &\tfrac{\partial}{\partial t}\sigma_1(x,t) + \tfrac{\partial}{\partial x}\sigma_1(x,t) \le C\sigma_1(x,t),\\ &\tfrac{\partial}{\partial t}\sigma_2(x,t) - \tfrac{\partial}{\partial x}\sigma_2(x,t) \le C\sigma_2(x,t),\\ &\tfrac{\partial}{\partial t}\sigma_3(x,t) \le \sigma_1(x,t)\sigma_2(x,t) - K\sigma_3(x,t), \end{aligned} \tag{3}$$

where C and K are constants and K is positive if $u_1^\epsilon, u_2^\epsilon$, and u_3^ϵ are all nonnegative. Thus, the sequence u_1^ϵ will be oscillatory only if the initial data $u_1^\epsilon(\cdot, 0)$ oscillates; in other words, oscillations of u_1^ϵ cannot be created by oscillations of sequences u_2^ϵ and u_3^ϵ. A similar statement is true for u_2^ϵ but not for u_3^ϵ. Assume that the initial data $u_3^\epsilon(\cdot, 0)$ converges strongly in L^2. The last inequality in (3) shows that oscillations in u_3^ϵ cannot be created if only one of the sequences u_1^ϵ and u_2^ϵ is oscillatory. However, one cannot rule out the possibility of creating oscillations if both u_1^ϵ and u_2^ϵ oscillate. In fact, the creation of oscillations in the sequence u_3^ϵ can be demonstrated by looking at initial data in a periodically modulated form (see McLaughlin, Papanicolaou, and Tartar [4]).

Our goal is to analyze the full 2D Broadwell model. However, as we will see, this cannot be done without introducing new tools (such as H-measures). In this note we consider a sequence $u^\epsilon = (u_1^\epsilon, u_2^\epsilon, u_3^\epsilon, u_4^\epsilon) \in L^\infty \cap L^2$ of solutions to the Broadwell model

$$\begin{aligned} &\tfrac{\partial}{\partial t}u_1^\epsilon + \tfrac{\partial}{\partial x}u_1^\epsilon = u_3^\epsilon u_4^\epsilon - u_1^\epsilon u_2^\epsilon,\\ &\tfrac{\partial}{\partial t}u_2^\epsilon - \tfrac{\partial}{\partial x}u_2^\epsilon = u_3^\epsilon u_4^\epsilon - u_1^\epsilon u_2^\epsilon,\\ &\tfrac{\partial}{\partial t}u_3^\epsilon + \tfrac{\partial}{\partial y}u_3^\epsilon = u_1^\epsilon u_2^\epsilon - u_3^\epsilon u_4^\epsilon,\\ &\tfrac{\partial}{\partial t}u_4^\epsilon - \tfrac{\partial}{\partial y}u_4^\epsilon = u_1^\epsilon u_2^\epsilon - u_3^\epsilon u_4^\epsilon \end{aligned} \tag{4}$$

defined on a strip $\boldsymbol{R}^2 \times [0,T] = \{(x,y,t) : 0 \le t < T\}$ and satisfying the initial conditions

$$(5)\ (u_1^\epsilon, u_2^\epsilon, u_3^\epsilon, u_4^\epsilon)(x,y,0) = (u_{01}^\epsilon, u_{02}^\epsilon, u_{03}^\epsilon, u_{04}^\epsilon)(x,y), \qquad u_{0i}^\epsilon \in L^\infty(\boldsymbol{R}^2) \cap L^2(\boldsymbol{R}^2).$$

We assume that

$$(6) \qquad \|u_i^\epsilon\|_{L^\infty(R^2\times[0,T])} \le M, \qquad i = 1,2,3,4$$

and note that local existence results and local bounds of the type (6) are easily obtained for small time intervals $[0,t]$, provided that the initial data $(u_{01}^\epsilon, u_{02}^\epsilon, u_{03}^\epsilon, u_{04}^\epsilon)$ lies in a bounded set of L^∞.

It is relatively simple to show that if the sequence $u_0^\epsilon = (u_{01}^\epsilon, u_{02}^\epsilon, u_{03}^\epsilon, u_{04}^\epsilon)$ converges strongly in $L^2(\boldsymbol{R}^2)$ to $u_0^0 = (u_{01}^0, u_{02}^0, u_{03}^0, u_{04}^0)$, then the solutions $u^\epsilon = (u_1^\epsilon, u_2^\epsilon, u_3^\epsilon, u_4^\epsilon)$ converge strongly in $[L^2_{loc}(\boldsymbol{R}^2 \times [0,T])]^4$ to a function $u^0 = (u_1^0, u_2^0, u_3^0, u_4^0)$,

which is a solution of (4). This fact was exploited by Peszek in [5] to show a strong convergence of certain finite difference approximations to systems of the type (1) (also see Peszek [6]).

We ask the following question: *what happens if exactly one of the sequences u_{0i}^ϵ (say u_{01}^ϵ) oscillates and if the others converge strongly in L^2?* Will u_2^0, u_3^0, or u_4^0 oscillate at later times or will they converge strongly? This question cannot be answered using Young measures simply because the answer depends on the direction in which the initial data oscillates and the Young measures do not carry such information. We will give an answer to this question with the help of H-measures.

We distinguish three directions in the (x, y)-plane given by unit vectors

$$v_1 = (0, 1), \quad v_2 = (\sqrt{2}/2, \sqrt{2}/2), \quad v_3 = (\sqrt{2}/2, -\sqrt{2}/2).$$

It turns out that u_2^0, u_3^0, and u_4^0 will converge strongly if u_{01}^ϵ does not oscillate in either of these directions (that is, if the H-measure μ associated with u_{01}^ϵ has no Dirac masses at the direction points $\pm v_1, \pm v_2$, and $\pm v_3$). This result will be proven in §3.

Conversely, one can construct examples of u_{01}^ϵ oscillating in the v_1-direction and such that oscillations of amplitude $O(t)$ are created in u_2^ϵ. Similarly, if u_{01}^ϵ oscillates in the direction of v_2 (or in the direction of v_3), then $O(t)$ oscillations will, as some examples show, be generated in u_3^ϵ (or in u_4^ϵ). We address this topic in more detail in §4.

We must point out that the described result extends easily to the situation in which both u_{01}^ϵ and u_{03}^ϵ (or both u_{01}^ϵ and u_{04}^ϵ) are oscillatory, provided that oscillations of u_{01}^ϵ do not occur in either of the specified directions, v_1, v_2, and v_3, and that oscillations of u_{03}^ϵ (or u_{04}^ϵ) do not occur in the directions $\pm v_2, \pm v_3$, and $\pm(0, 1)$ (see the end of §4 for a detailed discussion).

2. Basic facts about H-measures. In this section we introduce H-measures and review their basic properties. We refer the reader to the original paper of Tartar [1] for all proofs and for a more detailed exposition.

For the sake of generality, we consider a sequence of vector valued functions $V^\epsilon : \Omega \to \boldsymbol{R}^p$ defined on an open set $\Omega \subset \boldsymbol{R}^N$. We assume that V^ϵ converges weakly in $(L^2(\Omega))^p$ to a function $\overline{V} \in (L^2(\Omega))^p$. We define S^{N-1},

$$S^{N-1} = \{\xi \in \boldsymbol{R}^N : \xi_1^2 + \cdots + \xi_N^2 = 1\},$$

to be the set of direction points and put $U^\epsilon = V^\epsilon - \overline{V}$. It can be shown that after extracting a subsequence (for which we will still keep the index ϵ), there exists a family $\mu = \mu^{ij}$ of complex-valued Radon measures, $\operatorname{supp}(\mu^{ij}) \subset \Omega \times S^{N-1}$, such that

$$\langle \mu^{ij}, \phi_1 \phi_2{}^* \otimes \psi \rangle = \lim_{\epsilon \to 0} \int_{R^N} \mathcal{F}(\phi_1 U_i^\epsilon)(\zeta) [\mathcal{F}(\phi_2 U_j^\epsilon)(\zeta)]^* \psi(\zeta/|\zeta|) d\zeta \tag{7}$$

$$\forall \phi_1, \phi_2 \in C_0(\Omega), \psi \in C(S^{N-1}),$$

where $\mathcal{F}$ denotes the Fourier transform and z^* denotes the complex conjugate of z. The family μ is called the H-measure associated with the extracted subsequence V^ϵ. It turns out that μ measures the oscillation and concentration effects. In particular, $\mu^{ij} = 0$ for all $i, j = 1, \ldots, p$, corresponds to strong convergence (in L^2_{loc}) of the extracted subsequence.

H-measures are Hermitian and nonnegative; that is,

$$\mu^{ij} = \mu^{ji}, \quad i,j = 1,\dots,p, \quad \text{and} \quad \sum_{i,j=1}^{p} \mu^{ij}\phi_i\phi_j^* \geq 0$$

for every $\phi_1,\dots\phi_p \in C_0(\Omega)$, in particular

$$\operatorname{supp}(\mu^{ij}) \subset \operatorname{supp}(\mu^{ii}) \cap \operatorname{supp}(\mu^{jj}), \qquad i,j = 1,\dots,p.$$

One obtains the following localization principle (which corresponds to the usual compensated compactness theorem).

If V^ϵ satisfy

$$\sum_{i=1}^{p}\sum_{k=1}^{N} \frac{\partial}{\partial x_k}[A_{jk}V_j^\epsilon] \to 0 \text{ strongly in } H_{loc}^{-1}(\Omega) \tag{8}$$

where $A_{jk} \in C(\Omega)$, then

$$\sum_{i=1}^{p}\sum_{k=1}^{N} A_{jk}\xi_k\mu^{jm} = 0, \qquad m = 1,\cdots,p \tag{9}$$

in $\Omega \times S^{N-1}$.

As a trivial example we consider a sequence $v^\epsilon : \Omega \to \boldsymbol{R}$ and let μ^v denote the H-measure associated with this sequence. We let $w^\epsilon : \Omega \times (0,1) \to \boldsymbol{R}$ be defined by

$$w^\epsilon(x_1,\dots,x_N,x_{N+1}) = v^\epsilon(x_1,\dots,x_N). \tag{10}$$

The localization principle applied to w^ϵ shows that the support of the H-measure μ^w associated with a subsequence of w^ϵ is included in $(\Omega \times (0,1)) \times S_0^{N-1}$, where $S_0^{N-1} \subset S^N$ is defined by

$$S_0^{N-1} = \{\xi \in S^N : \xi_{N+1} = 0\} \tag{11}$$

and can be identified, in a natural way, with S^{N-1}. We point out that a more careful analysis of (7) gives a relation between μ^w and μ^v. It can be shown that, for all $0 < a < b < 1$ and all measurable sets $A \subset \Omega$ and $B \subset S^{N-1}$ $(= S_0^{N-1})$,

$$\mu^w(A \times (a,b) \times B) = (b-a)\mu^v(A \times B). \tag{12}$$

Unlike the commonly used Young measures, the H-measures satisfy transport properties.

Assume that $V^\epsilon = (v_1^\epsilon, v_2^\epsilon)$ satisfy

$$\sum_{j=1}^{N} b_j \frac{\partial v_1^\epsilon}{\partial x_j} = v_2^\epsilon, \tag{13}$$

where b_j are constants. Then, the H-measure μ associated with $(v_1^\epsilon, v_2^\epsilon)$ satisfies

$$-\left\langle \mu^{11}, \sum_{j=1}^{N} b_j \frac{\partial}{\partial x_j}\phi \right\rangle = 2\langle \operatorname{Re}\, \mu^{12}, \phi\rangle \tag{14}$$

for every test function $\phi(x,\xi)$ of class C^1 on $\Omega \times S^{N-1}$ with compact support.

Finally, we note that one can obtain a formula for change of coordinates. Here we restrict our attention to the simplest case. We let $W^\epsilon(x) = V^\epsilon(F(x))$, where F is affine; $F(x) = Ax + b$ and A is an invertible constant matrix. We also let $\pi = \pi^{ij}$ denote the H-measure associated with W^ϵ. One obtains the following formula:

$$\langle \pi^{ij}, \phi \otimes \psi \rangle = \left\langle \mu^{ij}, \frac{1}{\det A} \phi(F^{-1}(x)) \otimes \psi\left(\frac{A^t\xi}{|A^t\xi|}\right) \right\rangle, \tag{15}$$

for every i and j.

3. Main result. We consider a sequence $u^\epsilon = (u_1^\epsilon, u_2^\epsilon, u_3^\epsilon, u_4^\epsilon) \in L^\infty \cap L^2$ of solutions to the Broadwell system (4). We assume that $u^\epsilon = (u_1^\epsilon, u_2^\epsilon, u_3^\epsilon, u_4^\epsilon)$ satisfy

$$||u_i^\epsilon||_{L^\infty(R^2 \times [0,T])} \leq M, \qquad i = 1,2,3,4 \tag{16}$$

and denote the initial data by

$$(u_1^\epsilon, u_2^\epsilon, u_3^\epsilon, u_4^\epsilon)(x,y,0) = (u_{01}^\epsilon, u_{02}^\epsilon, u_{03}^\epsilon, u_{04}^\epsilon)(x,y). \tag{17}$$

We assume that $u_{02}^\epsilon, u_{03}^\epsilon$, and u_{04}^ϵ converge strongly in L^2 to u_{02}^0, u_{03}^0, and u_{04}^0, respectively, and that u_{01}^ϵ converges weakly in L^2 to $\overline{u}_{01}$. We also assume that the sequence u_{01}^ϵ is associated with the H-measure μ.

In this section we will prove the following result.

THEOREM 3.1. *Assume that μ has no Dirac masses at the direction points $\pm v_1$, $\pm v_2$ and $\pm v_3$ defined by*

$$v_1 = (0,1), \quad v_2 = (\sqrt{2}/2, \sqrt{2}/2), \quad v_3 = (\sqrt{2}/2, -\sqrt{2}/2).$$

Then $u_2^\epsilon, u_3^\epsilon$, and u_4^ϵ converge strongly in $L^2_{loc}(\boldsymbol{R}^2 \times [0,T])$ to u_2^0, u_3^0, and u_4^0, respectively, the sequence u_1^ϵ converges weakly to $\overline{u}_1$, and the function $u^0 = (\overline{u}_1, u_2^0, u_3^0, u_4^0)$ is the solution of (1) with the initial conditions

$$u^0(x,y,0) = (\overline{u}_{01}, u_{02}^0, u_{03}^0, u_{04}^0)(x,y).$$

Proof. It is enough to prove that $u_2^\epsilon, u_3^\epsilon$, and u_4^ϵ converge strongly. One can easily show that there is at most one solution $u^0 = (\overline{u}_1, u_2^0, u_3^0, u_4^0) \in (L^\infty(\boldsymbol{R}^2 \times [0,T)))^4$ of (1) that satisfies initial conditions $u^0(x,y,0) = (\overline{u}_{01}, u_{02}^0, u_{03}^0, u_{04}^0)(x,y)$. Thus, it is sufficient to show that there exists a subsequence of $(u_2^\epsilon, u_3^\epsilon, u_4^\epsilon)$ (for which we will still keep the index ϵ) that converges strongly to (u_2^0, u_3^0, u_4^0).

One easily obtains the following estimates:

$$\begin{aligned} u_1^\epsilon(x,y,t) &= u_{01}^\epsilon(x-t,y) + O(t), \\ u_2^\epsilon(x,y,t) &= u_{02}^\epsilon(x+t,y) + O(t), \\ u_3^\epsilon(x,y,t) &= u_{03}^\epsilon(x,y-t) + O(t), \\ u_4^\epsilon(x,y,t) &= u_{04}^\epsilon(x,y+t) + O(t), \end{aligned} \tag{18}$$

where the terms $O(t)$ are bounded independently of ϵ (by a function of order $O(t)$).

Plugging this into (4) yields

$$
\begin{aligned}
\tfrac{\partial}{\partial t}u_1^\epsilon(x,y,t)+\tfrac{\partial}{\partial x}u_1^\epsilon(x,y,t) &= u_{03}^\epsilon(x,y-t)u_{04}^\epsilon(x,y+t)\\
&\quad -u_{01}^\epsilon(x-t,y)u_{02}^\epsilon(x+t,y)+O(t),\\
\tfrac{\partial}{\partial t}u_2^\epsilon(x,y,t)-\tfrac{\partial}{\partial x}u_2^\epsilon(x,y,t) &= u_{03}^\epsilon(x,y-t)u_{04}^\epsilon(x,y+t)\\
&\quad -u_{01}^\epsilon(x-t,y)u_{02}^\epsilon(x+t,y)+O(t),\\
\tfrac{\partial}{\partial t}u_3^\epsilon(x,y,t)+\tfrac{\partial}{\partial y}u_3^\epsilon(x,y,t) &= u_{01}^\epsilon(x-t,y)u_{02}^\epsilon(x+t,y)\\
&\quad -u_{03}^\epsilon(x,y-t)u_{04}^\epsilon(x,y+t)+O(t),\\
\tfrac{\partial}{\partial t}u_4^\epsilon(x,y,t)-\tfrac{\partial}{\partial y}u_4^\epsilon(x,y,t) &= u_{01}^\epsilon(x-t,y)u_{02}^\epsilon(x+t,y)\\
&\quad -u_{03}^\epsilon(x,y-t)u_{04}^\epsilon(x,y+t)+O(t),
\end{aligned}
\tag{19}
$$

and

$$
\begin{aligned}
u_1^\epsilon(x,y,t) &= u_{01}^\epsilon(x-t,y)\left(1-\textstyle\int_0^t u_{02}^\epsilon(x-t+2s,y)ds\right)\\
&\quad +\textstyle\int_0^t u_{03}^\epsilon(x-t+s,y-s)u_{04}^\epsilon(x-t+s,y+s)ds+O(t^2),\\
u_2^\epsilon(x,y,t) &= u_{02}^\epsilon(x+t,y)\left(1-\textstyle\int_0^t u_{01}^\epsilon(x+t-2s,y)ds\right)\\
&\quad +\textstyle\int_0^t u_{03}^\epsilon(x+t-s,y-s)u_{04}^\epsilon(x+t-s,y+s)ds+O(t^2),\\
u_3^\epsilon(x,y,t) &= u_{03}^\epsilon(x,y-t)\left(1-\textstyle\int_0^t u_{04}^\epsilon(x,y-t+2s)ds\right)\\
&\quad +\textstyle\int_0^t u_{01}^\epsilon(x-s,y-t+s)u_{02}^\epsilon(x+s,y-t+s)ds+O(t^2),\\
u_4^\epsilon(x,y,t) &= u_{04}^\epsilon(x,y+t)\left(1-\textstyle\int_0^t u_{03}^\epsilon(x,y+t-2s)ds\right)\\
&\quad +\textstyle\int_0^t u_{01}^\epsilon(x-s,y+t-s)u_{02}^\epsilon(x+s,y+t-s)ds+O(t^2),
\end{aligned}
\tag{20}
$$

with the terms $O(t)$ and $O(t^2)$ bounded independently of ϵ (by functions of order $O(t)$ and $O(t^2)$, respectively).

We define the sequence of approximate solutions $u^{\epsilon,h}$ in a recursive way by requiring that for all $k=1,2,\dots,[T/h]$ and $0\le t\le h$

$$
\begin{aligned}
u_1^{\epsilon,h}(x,y,kh+t) &= u_1^{\epsilon,h}(x-t,y,kh)\left(1-\textstyle\int_0^t u_2^{\epsilon,h}(x-t+2s,y,kh)ds\right)\\
&\quad +\textstyle\int_0^t u_3^{\epsilon,h}(x-t+s,y-s,kh)u_4^{\epsilon,h}(x-t+s,y+s,kh)ds,\\
u_2^{\epsilon,h}(x,y,kh+t) &= u_2^{\epsilon,h}(x+t,y,kh)\left(1-\textstyle\int_0^t u_1^{\epsilon,h}(x+t-2s,y,kh)ds\right)\\
&\quad +\textstyle\int_0^t u_3^{\epsilon,h}(x+t-s,y-s,kh)u_4^{\epsilon,h}(x+t-s,y+s,kh)ds,\\
u_3^{\epsilon,h}(x,y,kh+t) &= u_3^{\epsilon,h}(x,y-t,kh)\left(1-\textstyle\int_0^t u_4^{\epsilon,h}(x,y-t+2s,kh)ds\right)\\
&\quad +\textstyle\int_0^t u_1^{\epsilon,h}(x-s,y-t+s,kh)u_2^{\epsilon,h}(x+s,y-t+s,kh)ds,\\
u_4^{\epsilon,h}(x,y,kh+t) &= u_4^{\epsilon,h}(x,y+t,kh)\left(1-\textstyle\int_0^t u_3^{\epsilon,h}(x,y+t-2s,kh)ds\right)\\
&\quad +\textstyle\int_0^t u_1^{\epsilon,h}(x-s,y+t-s,kh)u_2^{\epsilon,h}(x+s,y+t-s,kh)ds,
\end{aligned}
\tag{21}
$$

and that $u_i^{\epsilon,h}(x,y,0)=u_{0i}^\epsilon(x,y)$ for $i=1,2,3,4$.

Obviously (20) yields that

$$
u_i^{\epsilon,h}(x,y,t)-u_i^\epsilon(x,y,t)=O(t^2)\text{ for }0\le t\le h\quad\text{and}\quad i=1,2,3,4.
$$

One may apply an inductive argument to show that for all $k=1,2,\dots,[T/h]$, and $0\le t\le h$,

$$
u_i^{\epsilon,h}(x,y,kh+t)-u_i^\epsilon(x,y,kh+t)=O((kh+t)h)
\tag{22}
$$

and, thus, that

$$||u_i^{\epsilon,h} - u_i^{\epsilon}||_{L^\infty(R^2\times[0,T])} = O(Th), \qquad i = 1,2,3,4. \tag{23}$$

This follows from applying estimates of the type (20) to consecutive time intervals $[kh,(k+1)h]$ with the initial conditions $u_{0i}^{\epsilon}(x,y)$ replaced by

$$u_i^{\epsilon}(x,y,kh) = u_i^{\epsilon,h}(x,y,kh) + O(kh^2), \qquad i = 1,2,3,4.$$

One easily obtains that, for sufficiently small h and $0 \le t \le h$,

$$u_i^{\epsilon}(x,y,kh+t) = u_i^{\epsilon,h}(x,y,kh+t) + O((kh+t)h), \qquad i = 1,2,3,4.$$

We observe that the terms $O(Th)$ in (23) are bounded independently of ϵ (by a function of order $O(Th)$).

We proceed in a recursive way. For each k we extract (if necessary) a subsequence of $u^{\epsilon,h}(\cdot,\cdot,kh)$, $\epsilon \to 0$ from the subsequence $\epsilon \to 0$ extracted in the previous step $k-1$. We denote $\overline{\mu}_{kh}$ to be the H-measures associated with the subsequences $u_1^{\epsilon,h}(\cdot,\cdot,kh)$. From (23) it suffices to show that extracted subsequences $u_i^{\epsilon,h}$ converge strongly for every given h and $i = 2,3,4$. This, in turn, is equivalent to showing that, for $k = 1,2,\cdots[T/h]$,

(I) if $\overline{\mu}_{kh}$ has no Dirac masses at points $\pm v_1, \pm v_2$, and $\pm v_3$, then $\overline{\mu}_{(k+1)h}$ has no Dirac masses at these points,

(II) extracted subsequences $u_i^{\epsilon,h}$, $i = 2,3,4$ converge strongly in $L^2_{loc}(\boldsymbol{R}^2 \times [kh,(k+1)h])$.

We note that part (I) and part (II) can be applied in a recursive way to consecutive time intervals to show the claimed convergence and that it is sufficient to show part (I) and part (II) for $k = 0$.

Part (I) follows from part (II) since

$$\begin{aligned} u_1^{\epsilon,h}(x,y,h) &= u_1^{\epsilon,h}(x-h,y,0)\left(1 - \textstyle\int_0^h u_2^{\epsilon,h}(x-h+2s,y,0)ds\right) \\ &\quad + \textstyle\int_0^h u_3^{\epsilon,h}(x-h+s,y-s,0)u_4^{\epsilon,h}(x-h+s,y+s,0)ds \\ &= u_{01}^{\epsilon}(x-h,y)A^{\epsilon,h}(x,y) + B^{\epsilon,h}(x,y), \end{aligned}$$

and since $A^{\epsilon,h}$ and $B^{\epsilon,h}$ converge strongly as $\epsilon \to 0$ (which follows from part (I)).

The proof of Part (II) follows directly from localization and transport results for H-measures. We will show that oscillations are not created in $u_3^{\epsilon,h}$ for $t \in [0,h]$. Proofs for $u_2^{\epsilon,h}$ and $u_4^{\epsilon,h}$ are similar and are therefore omitted.

From (21) we obtain that, for $t \in [0,h]$,

$$\begin{aligned} u_3^{\epsilon,h}(x,y,t) &= u_{03}^{\epsilon}(x,y-t)\left(1 - \textstyle\int_0^t u_{04}^{\epsilon}(x,y-t+2s)ds\right) \\ &\quad + \textstyle\int_0^t u_{01}^{\epsilon}(x-s,y-t+s)u_{02}^{\epsilon}(x+s,y-t+s)ds \end{aligned}$$

and since $u_{02}^{\epsilon}, u_{03}^{\epsilon}$, and u_{04}^{ϵ} converge strongly we only need to show that

$$D^{\epsilon}(x,y,t) = \int_0^t u_{01}^{\epsilon}(x-s,y-t+s)u_{02}^{0}(x+s,y-t+s)ds \tag{24}$$

converges strongly in $L^2_{loc}(\boldsymbol{R}^2 \times [0,h])$. We will proceed in two steps. First, we will use the assumption that the H-measure μ associated with the sequence u_{01}^{ϵ} has no

Dirac masses at the direction points $\pm v_2 = \pm(\sqrt{2}/2, \sqrt{2}/2)$ to show the convergence of D^ϵ in the case of $u_{02}^0 \in C(\boldsymbol{R}^2)$.

We write (24) in the form

$$\frac{\partial}{\partial t} D^\epsilon(x,y,t) + \frac{\partial}{\partial y} D^\epsilon(x,y,t) = u_{01}^\epsilon(x-t,y) u_{02}^0(x+t,y), \tag{25}$$

let $U^\epsilon(x,y,t) = u_{01}^\epsilon(x-t,y)u_{02}^0(x+t,y)$, extract (if necessary) a subsequence of (D^ϵ, U^ϵ), and define $\nu = (\nu^{ij})$, $i,j = 1,2$, to be the H-measure associated with this subsequence, $\operatorname{supp}(\nu^{ij}) \subset \boldsymbol{R}^2 \times [0,h] \times S^2$. It can be easily shown that

$$\nu^{22} = |u_{02}^0(x+t,y)|^2 \overline{\nu},$$

where $\overline{\nu}$ is the H-measure associated with some subsequence of

$$\overline{U}^\epsilon(x,y,t) = u_{01}^\epsilon(x-t,y).$$

The localization principle yields that

$$\operatorname{supp}(\nu^{11}) \subset \big(\boldsymbol{R}^2 \times [0,h]\big) \times \{\xi_t + \xi_y = 0\}, \quad \operatorname{supp}(\nu^{22}) \subset \big(\boldsymbol{R}^2 \times [0,h]\big) \times \{\xi_t + \xi_x = 0\}$$

(where $(\xi_x, \xi_y, \xi_t) \in S^2$), and

$$\operatorname{supp}(\nu^{12}) \subset \big(\boldsymbol{R}^2 \times [0,h]\big) \times \{\pm(\sqrt{3}/3, \sqrt{3}/3, -\sqrt{3}/3)\}. \tag{26}$$

A more careful analysis shows that the H-measure $\overline{\nu}$ can be obtained from the H-measure μ associated with the sequence u_{01}^ϵ. To see that we recall that μ extends in a natural way to the H-measure μ^w associated with the sequence $w^\epsilon(x,y,t) = u_{01}^\epsilon(x,y)$ and that this extension is given by (11)–(12). H-measure $\overline{\nu}$ is then obtained from μ^w by making the change of variables $F(x,y,t) = (x-t, y, x+t)$ and applying the formula (15). We only point out that, since μ has no Dirac masses at the direction points $\pm v_2 = \pm(\sqrt{2}/2, \sqrt{2}/2)$, one obtains that measures $\overline{\nu}$ and ν^{22} have no Dirac masses at the points $\pm(\sqrt{3}/3, \sqrt{3}/3, -\sqrt{3}/3)$ and that the inclusion

$$\operatorname{supp}(\nu^{12}) \subset \operatorname{supp}(\nu^{11}) \cap \operatorname{supp}(\nu^{22}),$$

together with (26), yields $\nu^{12} \equiv 0$.

The transport property for H-measure ν gives the following equation:

$$\frac{\partial}{\partial t}\nu^{11} + \frac{\partial}{\partial y}\nu^{11} = 2\operatorname{Re}\, \nu^{12} = 0 \tag{27}$$

((27) holds in the distributional sense). Since $|D^\epsilon(x,y,t)| \le Ct$ independently of ϵ we obtain that $\nu^{11} \equiv 0$, and that the extracted subsequence D^ϵ converges strongly in $L^2_{loc}(\boldsymbol{R}^2 \times [0,h])$.

Now we can consider the general case of $u_{02}^0 \in L^\infty(\boldsymbol{R}^2) \cap L^2(\boldsymbol{R}^2)$. We observe that if $\hat{u}_{02}^h$ denotes a sequence of continuous functions converging strongly in L^2 to u_{02}^0 and

$$D^{\epsilon,h}(x,y,t) = \int_0^t u_{01}^\epsilon(x-s, y-t+s)\hat{u}_{02}^h(x+s, y-t+s)ds, \tag{28}$$

then

$$||D^{\epsilon,h} - D^\epsilon||_{L^2(R^2\times[0,h])} \le C||\hat{u}_{02}^0 - u_{02}^h||_{L^2(R^2)}. \tag{29}$$

The previous argument shows that $D^{\epsilon,h}$ converges in L^2_{loc} for every given h, and, therefore, from (29) it follows that D^ϵ converges strongly in $L^2_{loc}(\boldsymbol{R}^2 \times [0,h])$.

This observation completes the proof of the theorem. □

4. Concluding remarks. Our concern is in studying sequences $(u_1^\epsilon, u_2^\epsilon, u_3^\epsilon, u_4^\epsilon)$ of solutions to the Broadwell model (4). We ask the question: What are the conditions that guarantee that the oscillatory sequence of initial conditions u_{01}^ϵ will produce oscillations in sequences $u_2^\epsilon, u_3^\epsilon$, and u_4^ϵ for later times t? The theorem proven in the previous section shows that $u_2^\epsilon, u_3^\epsilon$, and u_4^ϵ will not oscillate if the H-measure μ associated with u_{01}^ϵ has no Dirac masses at the direction points

$$v_1 = (0,1), \quad v_2 = (\sqrt{2}/2, \sqrt{2}/2), \quad \text{and} \quad v_3 = (\sqrt{2}/2, -\sqrt{2}/2).$$

One may conjecture that, if u_{01}^ϵ oscillates along the direction of the y-axis, then the integral term

$$\int_0^t u_{01}^\epsilon(x+t-2s, y)ds$$

will oscillate along the same direction and, thus, that oscillations of amplitude $O(t)$ will be created in u_2^ϵ (cf. (20)). Similarly, one may argue that if u_{01}^ϵ oscillates along the direction of v_2 (or along the direction of v_3), then the term

$$\int_0^t u_{01}^\epsilon(x-s, y-t+s)u_{02}^\epsilon(x+s, y-t+s)ds$$

$$\text{(or the term } \int_0^t \int_0^t u_{01}^\epsilon(x-s, y+t-s)u_{02}^\epsilon(x+s, y+t-s)ds \text{)}$$

will oscillate, creating $O(t)$ oscillations of u_3^ϵ (or of u_4^ϵ).

These speculations can be supported by considering u_{01}^ϵ in the form

$$u_{01}^\epsilon(x,y) = \sum_j w_j(x,y)e^{i(xm_1^j/\epsilon + ym_2^j/\epsilon)}$$

and by noting that

$$\int_0^t u_{01}^\epsilon(x+t-2s, y)ds = \sum_{\{j:m_1^j \neq 0\}} e^{iym_2^j/\epsilon} \int_0^t w_j(x+t-2s, y)e^{i(x+t-2s)m_1^j/\epsilon}ds$$
$$+ \sum_{\{j:m_1^j=0\}} e^{iym_2^j/\epsilon} \int_0^t w_j(x+t-2s, y)ds.$$

We observe that the first sum in the above equation converges strongly to zero, since the sequence $e^{2ism_1^j/\epsilon}$ converges weakly to 0 if $m_1^j \neq 0$, and that the second sum converges only weakly. This observation, together with (20), guarantees that oscillations of magnitude $O(t)$ are created in the sequence u_2^ϵ. We also point out that the same type of arguments can be applied to support the above conjecture about creation of oscillations in sequences u_3^ϵ and u_4^ϵ.

It turns out that we cannot exhibit the phenomenon of creation of oscillations in u_2^ϵ, u_3^ϵ, and u_4^ϵ by using the language of H-measures. To illustrate this we will construct a sequence u_{01}^ϵ with the H-measure having Dirac masses at points $\pm v_1$ and such that the $O(t)$ amplitude oscillations are *not* created in u_2^ϵ.

To do this we fix t and denote by ν the H-measure associated with a subsequence of $U^{\epsilon,t}(x,y) = \int_0^t u_{01}^\epsilon(x+t-2s,y)ds$. We write

$$\int_0^t u_{01}^\epsilon(x+t-2s,y)ds = \frac{1}{2}\left(u_{01}^\epsilon(\cdot,y) \overset{*}{\star} \chi_{[-t,t]}\right)(x)$$

and observe that, for $\phi_1, \phi_2 \in C_{00}^1(\boldsymbol{R}^2)$ and $\psi \in C(S^1)$,

$$\begin{aligned}\langle \nu, \phi_1\phi_2 \times \psi\rangle &= \lim_{\epsilon\to 0}\frac{1}{4}\int_{R^2}\mathcal{F}(\phi_1(u_{01}^\epsilon \star \chi_{[-t,t]}))(\zeta)[\mathcal{F}(\phi_2(u_{01}^\epsilon \star \chi_{[-t,t]}))(\zeta)]^*\psi(\zeta/|\zeta|)d\zeta \\ &= \lim_{\epsilon\to 0}\frac{1}{4}\int_{R^2}\mathcal{F}((\phi_1 u_{01}^\epsilon)\star\chi_{[-t,t]})(\zeta)[\mathcal{F}((\phi_2 u_{01}^\epsilon)\star\chi_{[-t,t]})(\zeta)]^*\psi(\zeta/|\zeta|)d\zeta + O(t^3) \\ &= \lim_{\epsilon\to 0}\int_{R^2}\mathcal{F}(\phi_1 u_{01}^\epsilon)(\zeta)[\mathcal{F}(\phi_2 u_{01}^\epsilon)(\zeta)]^*\psi(\zeta/|\zeta|)\frac{\sin^2(t\zeta_x)}{|\zeta_x|^2}d\zeta + O(t^3). \qquad (30)\end{aligned}$$

One can construct a sequence u_{01}^ϵ in such a way that $\mathcal{F}(u_{01}^\epsilon)$ has its support contained in two balls moving to infinity along the curve $\zeta_y = \zeta_x^2$, $\zeta_x > 0$, and $\zeta_y = -\zeta_x^2$, $\zeta_x < 0$. One may take, for example,

$$u_{01}^\epsilon = w(x,y)\left(e^{i(x\epsilon^{-1}+y\epsilon^{-2})} + e^{-i(x\epsilon^{-1}+y\epsilon^{-2})}\right),$$

where w has bounded support. The H-measure μ associated with u_{01}^ϵ constructed above has two Dirac masses at direction points $\pm v_1$. The formula (30), on the other hand, implies that the H-measure ν associated with a subsequence of integral terms $U^{\epsilon,t}(x,y) = \int_0^t u_{01}^\epsilon(x+t-2s,y)ds$ is zero. This observation, together with the estimate shown in the second equation in (20), guarantees that there are no oscillatory terms of magnitude $O(t)$ in u_2^ϵ.

Finally, we would like to address the problem in which oscillations are imposed initially in more than one of the sequences u_{01}^ϵ,u_{02}^ϵ, u_{03}^ϵ, and u_{04}^ϵ. The theorem proven in the previous section extends easily to the situations in which u_{01}^ϵ and u_{03}^ϵ or u_{01}^ϵ and u_{04}^ϵ are oscillatory, provided that oscillations of u_{01}^ϵ do not occur in either of the specified directions and that oscillations of u_{03}^ϵ (or u_{04}^ϵ) do not occur in the directions $\pm v_2, \pm v_3$, and $\pm(0,1)$. Similarly, one may deal with the cases in which u_{02}^ϵ and either u_{03}^ϵ or u_{04}^ϵ oscillate.

It is much more difficult to characterize the cases in which both u_{01}^ϵ and u_{02}^ϵ oscillate (or both u_{03}^ϵ and u_{04}^ϵ). It turns out that these problems cannot be solved if we use only H-measures. To illustrate this we consider the term

$$\int_0^t u_{01}^\epsilon(x-s,y-t+s)u_{02}^\epsilon(x+s,y-t+s)ds \qquad (31)$$

occurring in the third equation in (20). We let

$$\begin{aligned}u_{01}^\epsilon(x,y) &= w(x,y)e^{i(xm_1^\epsilon+ym_2^\epsilon)}, \\ u_{02}^\epsilon(x,y) &= v(x,y)e^{i(xn_1^\epsilon+yn_2^\epsilon)},\end{aligned}$$

and observe that (31) takes the form

$$e^{i(x(m_1^\epsilon+n_1^\epsilon)+(y-t)(m_2^\epsilon+n_2^\epsilon))}\int_0^t w(x-s,y-t+s)v(x+s,y-t+s)e^{i(-m_1^\epsilon+m_2^\epsilon+n_1^\epsilon+n_2^\epsilon)s}ds.$$

Thus, the creation of oscillations depends on the behavior of $-m_1^\epsilon + m_2^\epsilon + n_1^\epsilon + n_2^\epsilon$. One needs, therefore, to compare not only the directions in which u_{01}^ϵ and u_{02}^ϵ oscillate but also the relative frequencies of their oscillations. This suggest the use of semiclassical measures of Gerard [8] or Wigner measures of Lions and Paul [9]. We hope to address this problem in the future.

Acknowledgment. The ideas presented in this paper originated from many fruitful discussions between the author and Dr. Luc Tartar. The author thanks Dr. Luc Tartar for introducing him to many new techniques and for his encouragement.

REFERENCES

[1] L. TARTAR, *H-measures, a new approach for studying homogenization, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 115 A (1990), pp. 193–230.

[2] ———, *Oscillations for semi-linear hyperbolic systems*, Postdoc/Visitor Seminar Notes, Sept. 8, 1992, Carnegie Mellon University, Pittsburgh, PA.

[3] ———, *Oscillations and asymptotic behavior for two semilinear hyperbolic systems*, in Dynamics of Infinite Dimensional Systems, NATO Adv. Sci. Inst. Ser. F, Comput. Systems Sci., 37, Springer-Verlag, Berlin, New York, 1987, pp. 341–356.

[4] D. W. MCLAUGHLIN, G. PAPANICOLAOU, AND L. TARTAR, *Weak limits of semilinear hyperbolic systems with oscillating data*, in Macroscopic Modeling of Turbulent Flows, Lecture Notes in Phys., Vol. 230, Springer-Verlag, Berlin, New York, 1985, pp. 277–289.

[5] R. PESZEK, *Young Measures and Convergence of Numerical Solutions to Systems of Semilinear Hyperbolic Equations*, manuscript, 1993.

[6] ———, *Instability and stability of numerical approximations to discrete velocity models of the Boltzmann equation*, Quart. Appl. Math., to appear.

[7] T. PŁATKOWSKI AND R. ILLNER, *Discrete velocity models of the Boltzmann equation: A survey of mathematical aspects of the theory*, SIAM Rev., 30 (1988), pp. 213–255.

[8] P. GERARD, *Mesures semi-classiques et ondes de Bloch*, Séminaire EDP 1990–1991, Ecole Polytechnique, Palaiseau, 1991.

[9] P. L. LIONS AND T. PAUL, *Sur les mesures de Wigner*, manuscript.

SIAM J. MATH. ANAL.
Vol. 26, No. 3, pp. 761–773, May 1995

014

FAMILIES OF TWO-POINT PADÉ APPROXIMANTS AND SOME ${}_4F_3(1)$ IDENTITIES *

JET WIMP† AND BERNHARD BECKERMANN‡

Abstract. In this paper, a family of two-point Padé approximants, that is, two polynomials, each of degree n and depending on integer $k, l, 0 \le l \le k \le n$, whose ratio approximates one function to order (z^{n+l+1}) at $z = 0$ and another to order $O(z^{-n+k})$ at $z = \infty$ is presented. The functions in question are ratios of Gaussian hypergeometric functions. Explicit closed-form expressions for the polynomials are given. Also, this derivation establishes some hypergeometric identities involving functions of the form ${}_4F_3(1)$. Several interesting limiting cases, namely, $[n, n]$ and $[n-1, n]$ Padé approximants for ratios of confluent hypergeometric functions and Bessel functions are given.

Key words. rational approximations, Padé approximations, hypergeometric functions, umbral calculus, confluent hypergeometric functions, Bessel functions, hypergeometric identities

AMS subject classifications. 33A30, 41A20, 41A21, 44A45

1. Introduction.

Many authors have discussed an unusual type of rational approximation. This approximation approximates one function to a certain order at $z = 0$ and another function to another order at $z = \infty$. For instance, in a recent paper [Hen3], Hendriksen showed that a rational approximant $P(z)/Q(z)$ had the property that P and Q were both of degree n and

$$\frac{P(z)}{Q(z)} = \begin{cases} \dfrac{z\,{}_2F_1\left(\begin{matrix}a, b+1\\ c+1\end{matrix}; z\right)}{{}_2F_1\left(\begin{matrix}a, b\\ c\end{matrix}; z\right)} + O(z^{n+1}), & z \to 0, \\ \dfrac{c\,{}_2F_1\left(\begin{matrix}b+1-c, b+1\\ b+2-a\end{matrix}; \dfrac{1}{z}\right)}{(b+1-a){}_2F_1\left(\begin{matrix}b+1-c, b\\ b+1-a\end{matrix}; \dfrac{1}{z}\right)} + O(z^{-n}), & z \to \infty. \end{cases} \tag{1.1}$$

This type of rational approximant is called a *two-point Padé approximant*; see [Jon2] for a reasonably complete bibliography. Actually, one may consider these approximations to be approximants to a special case of the two-point continued fractions discussed earlier by McCabe and Murphy [McC1], [McC2], and later by Cooper, Magnus, and McCabe [Coo]. Hendriksen was able to provide closed-form expressions for his approximants.

* Received by the editors December 12, 1991; accepted for publication (in revised form) November 22, 1993.

† Department of Mathematics and Computer Sciences, Drexel University, Philadelphia, Pennsylvania 19104, USA. The research of this author was supported by the National Science Foundation under Research Grant DMS 890 1610.

‡ Université des Sciences et Technologies de Lille, Laboratoire d'Analyse Numérique et d'Optimisation, UFR IEEA M3, 59655 Villeneuve d'Ascq Cédex, France.

In 1987 Wimp [Wim1] obtained a closed-form expression for a rational approximant $R(z)/S(z)$, R and S both of degree n, with the property

$$\frac{R(z)}{S(z)} = \frac{z\,{}_2F_1\left(\begin{matrix}a, b+1\\ c+1\end{matrix}; z\right)}{{}_2F_1\left(\begin{matrix}a, b\\ c\end{matrix}; z\right)} + O(z^{2n+1}), \qquad z \to 0. \tag{1.2}$$

$R(z)/S(z)$ is the $[n-1, n]$ Padé approximant to the function on the right, without the z factor. The formulas were closed-form expressions for rational approximants to the well-known continued fractions of Gauss; see [Kho, p. 133ff].

The above two formulas led us to suspect that one could find a *family* of Padé approximants with order terms transitional between the order terms in (1.1) and (1.2), respectively. Our suspicions turned out to be correct, and we present the derivation in this paper.

Hendriksen [Hen3] and Wimp [Wim1] presented explicit formulas for the polynomials they discussed. Our approach allows us to do the same. However, the previous authors used an unwieldly strategy that required the construction of bases of solutions of complicated difference equations. Our approach, based on the umbral calculus, more specifically, the use of projection operators applied to a two-element umbral calculus of formal series, avoids the difference equation approach. The same results may be obtained by the use of divided difference operators, but the umbral calculus approach is less cluttered and more clearly shows the way to generalization. (For some other recent applications of the umbral calculus, for instance, computing generalized Laplace integrals or determinants of Hessian type matrices with operator elements, see [Wim2]–[Wim5]. For background, we recommend the work of Garcia and Joni [Gar], and especially Roman and Rota, [Rom1]–[Rom3].)

Both Hendriksen and Wimp have commented on the prominent role that ${}_4F_3(1)$ identities play in the theory of Padé approximation and orthogonal polynomials. Hendriksen displayed a sophisticated ${}_4F_3(1)$ identity that arose in the theory of Jacobi–Laurent polynomials and asked whether it might be a special case of a known ${}_4F_3(1)$ identity. Wimp also derived several such identities. Our main results lead not only to classes of rational approximations but also to an abundance of ${}_4F_3(1)$ identities and reveals how these identities are related.

We point out that all of our recent discoveries in one-point Padé approximants are well known in the context of continued fractions; see, for example, the book of Wall [Wal], and in the context of T-fractions, some of the results for the two-point Padé cases have been obtained by Jones and Thron [Jon1, §7.3]. What have not been obtained by previous authors are closed-form expressions for the rational approximants of the relevant continued fractions.

For special functions in this paper we use the notation of the Bateman manuscript volumes [Erd]. For a background in Padé approximation, we recommend the references [Bak] and [Gil], and for the interface with continued fractions, the book [Jon1]. A very recent and extensive survey of the practical aspects of continued fractions, including two-point Padé approximants, is [Jon2]. This paper contains 149 references.

2. A little (umbral) calculus. We denote by $\mathcal{H}$ the linear space of formal

series with complex coefficients

$$\mathcal{H} = \left\{ h(z) \middle| h(z) = \sum_{j=-\infty}^{\infty} A_j z^j, A_j \in \mathcal{C} \right\}. \tag{2.1}$$

It is important to note that the indeterminant z *does not take on values.* It is best to think of z^j as simply a place marker in the series in (2.1). In fact, one could just as well write the elements of $\mathcal{H}$ as doubly infinite sequences $(\ldots, a_2, a_1, a_0, a_1, a_2, \ldots)$ with addition and scalar multiplication being defined in the obvious way.

Define the following projection operator onto a finite-dimensional subspace of $\mathcal{H}$:

$$\Pi_{r,s}^z\{f(z)\} = \begin{cases} \sum_{j=r}^{s} A_j z^j, & r \le s, \\ 0, & r > s. \end{cases} \tag{2.2}$$

This operator is called the (r, s) cut of h. We will often use the special cut operator $\Pi_{0,k}$. This is simply the kth partial sum of a "Taylor" series.

The following properties are easily verified:

$$\Pi_{r,s}^z\{f(z)\}|_{z\to 1/z} = z^{r-s}\Pi_{-r,s-2r}^z\{z^{s-r}f(1/z)\}, \tag{2.3}$$

$$\Pi_{r,s}^z\{z^p f(z)\} = z^p \Pi_{r-p,s-p}^z\{f(z)\}, \tag{2.4}$$

$$\Pi_{r,s}^z \Pi_{p,q}^z = \Pi_{k,l}^z, \quad k = \max\{r,p\}, \quad l = \min\{s,q\}. \tag{2.5}$$

Also of use will be the formal two-element series

$$h(z,w) = \sum_{i,j=-\infty}^{\infty} A_{i,j} z^i w^j. \tag{2.6}$$

For these series the operators (2.2) commute:

$$\Pi_{r,s}^z \Pi_{t,u}^w = \Pi_{t,u}^w \Pi_{r,s}^z. \tag{2.7}$$

Other useful properties of these operators are obvious and will be invoked as they are needed.

The subspace of $\mathcal{H}$ of those series containing only a finite number of negative powers, i.e., $A_j = 0$ except for a finite number of $j < 0$, is a field; call it $\mathcal{H}^+$. Multiplication is Cauchy multiplication of series and division is defined recursively by synthetic division. To multiply two such series of hypergeometric type, one uses the formula

$$\begin{aligned} {}_rF_s\left(\begin{matrix} a_r \\ b_s \end{matrix}; z\right) \times {}_mF_n\left(\begin{matrix} c_m \\ d_n \end{matrix}; z\right) \\ = \sum_{k=0}^{\infty} \frac{(c_m)_k z^k}{k!(d_n)_k} {}_{n+r+1}F_{s+m}\left(\begin{matrix} -k, a_r, 1-k-d_n \\ b_s, 1-k-c_m \end{matrix}; (-1)^{m+n+1}\right). \end{aligned} \tag{2.8}$$

(In the above, we employ the standard notation for representing expressions involving a string of parameters.)

We shall later require the formula for turning around a terminating hypergeometric function

$$(2.9)\qquad {}_{r+1}F_s\left(\begin{matrix}-k, a_r\\ b_s\end{matrix};z\right) = (-1)^k z^k \frac{(a_r)_k}{(b_s)_k}\, {}_{s+1}F_r\left(\begin{matrix}-k, 1-b_s-k\\ 1-a_r-k\end{matrix};\frac{(-1)^{r+s}}{z}\right).$$

Similarly, the subspace of $\mathcal{H}$ consisting of series with only a finite number of positive powers is a field; call it $\mathcal{H}^-$. When z is replaced by $1/z$ in (2.8) we get a similar formula in $\mathcal{H}^-$. (One reason why the umbral calculus is useful is that one can introduce an ingenious Laplace transform calculus on $\mathcal{H}^+$ and $\mathcal{H}^-$ in which concepts such as composition, fundamental shift theorems, and convolution have unexpected analogues; see [Rom2], [Wim2].)

There is a natural isomorphism between the field of rational complex functions and the subfield of rational functions in $\mathcal{H}^+$ or $\mathcal{H}^-$. Thus

$$(2.10)\qquad h(z) = \sum_{j=-\infty}^{m} A_j z^j, \qquad m \in \mathcal{Z},$$

is rational if and only if there is a linear difference operator acting on j with constant coefficients that annihilates the sequence $\{A_j\}_{-\infty}^{m}$. If that is the case, there is a complex rational function to which the above series converges for a complex variable z, $|z| > r$. Conversely, we may identify any such complex series with a series in $\mathcal{H}^-$ that represents a rational function. Analogous statements are true for series in $\mathcal{H}^+$.

One of our basic theorems in §4 involves an umbral calculus analogue of the ratio of gamma functions and its representation by means of Gaussian hypergeometric functions.

DEFINITION (umbral gamma functions).

$$(2.11)\qquad \begin{aligned}\Gamma_k(a,b,c;z) = z^k \exp\Bigg\{ \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m(m+1)z^m} &[\mathrm{B}_{m+1}(c) + \mathrm{B}_{m+1}(c-a-b)\\ &- \mathrm{B}_{m+1}(c-a) - \mathrm{B}_{m+1}(c-k-b)]\Bigg\}, \quad k \in \mathcal{Z},\end{aligned}$$

the series on the right being an element of $\mathcal{H}^-$ constructed in the obvious way. (The $\mathrm{B}_n(x)$ are the Bernoulli polynomials, in the notation of [Erd, §1.13].)

THEOREM 2.1. *For all series in* $\mathcal{H}^-$,

$$(2.12)\qquad {}_2F_1\left(\begin{matrix}a, b\\ c+z\end{matrix};1\right) = \Gamma_0(a,b,c;z),$$

$$(2.13)\qquad {}_2F_1\left(\begin{matrix}a, b\\ c-z\end{matrix};1\right) = \Gamma_0(a,-b,a+1-c;z).$$

Proof. The left-hand side of (2.12) is interpreted as a series in $\mathcal{H}^-$ constructed in the following way:

$$
\begin{aligned}
{}_2F_1\left(\begin{matrix}a,b\\c+z\end{matrix};1\right) &= \sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{(c+z)_k k!} = \sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{z^k k!}\sum_{m=0}^{\infty}\frac{\mu_m(c)}{z^m}\\
&= \sum_{m=0}^{\infty}\frac{1}{z^m}\sum_{k=0}^{m}\frac{(a)_k(b)_k\mu_{m-k}(c)}{k!} = \sum_{m=0}^{\infty}\frac{P_m(a)}{z^m},
\end{aligned}
$$

where $P_m(a)$ is a polynomial in a of degree m. (Note the second equality is simply the definition of $\mu_k(c)$.)

A little reflection convinces one that the right-hand side of (2.12), call it $V(a,z)$, can be written

$$
V(a,z) = \sum_{m=0}^{\infty}\frac{Q_m(a)}{z^m}, \tag{2.15}
$$

where $Q_m(a)$ is a polynomial in a of degree m. We shall show that $Q_m \equiv P_m$ by showing that they agree on the nonpositive integers.

Let $a = -p, p \in \mathcal{N}$. Using the fact that

$$
\mathrm{B}_{k+1}(x+1) - \mathrm{B}_{k+1}(x) = (k+1)x^k, \qquad k \in \mathcal{N}, \tag{2.16}
$$

allows us to write

$$
\frac{V(-p-1,z)}{V(-p,z)} = \frac{z+c+p-b}{z+c+p}, \tag{2.17}
$$

since the quantities in the exponents coming from (2.11) can be expressed in terms of logarithms. Using the fact that $V(0,z) = 1$ and iterating shows that

$$
V(-p,z) = \frac{(c+z-b)_p}{(c+z)_p}. \tag{2.18}
$$

Thus

$$
\begin{aligned}
Q_m(-p) &= \text{Coefficient } z^{-m} \text{ in } V(-p,z)\\
&= \text{Coefficient } z^{-m} \text{ in } \frac{(c+z-b)_p}{(c+z)_p}.
\end{aligned} \tag{2.19}
$$

However, in this case the left-hand side, by Vandermonde's theorem [Sla, p. 2], is the rational function $(c+z-b)_p/(c+z)_p$ when z is a complex variable, so by isomorphism the left-hand side represents the same rational function in $\mathcal{H}^-$. Thus $P_m(-p)$ is also given by the right-hand side above, so $Q_m = P_m$ on the nonpositive integers, hence the two are identical.

The second statement of the theorem follows when z is replaced by $-z$ and the relation

$$
\mathrm{B}_k(1-x) = (-1)^k\mathrm{B}_k(x), \qquad k \in \mathcal{N}, \tag{2.20}
$$

is used. □

The reader will observe that the exponential term in (2.11) is the asymptotic series for

$$\frac{\Gamma(z+c)\Gamma(z+c-a-b)}{\Gamma(z+c-a)\Gamma(z+c-k-b)}$$

when z is a real variable, $z \to \infty$. When either a or b is a negative integer, each series in Theorem 2.1 represents a rational function given by the terminating ${}_2F_1$ on the left or, equivalently, by the appropriate ratio of gamma functions rewritten in terms of Pochhammer symbols. One might be tempted to conclude that

$$(2.21)\qquad {}_2F_1\left(\begin{matrix}a,b\\c+z\end{matrix};1\right) = \frac{\Gamma(z+c)\Gamma(z+c-a-b)}{\Gamma(z+c-a)\Gamma(z+c-b)},$$

$$(2.22)\qquad {}_2F_1\left(\begin{matrix}a,b\\c-z\end{matrix};1\right) = \frac{\Gamma(z+a+1-c)\Gamma(z+b+1-c)}{\Gamma(z+1-c)\Gamma(z+a+b+1-c)}.$$

These equations are correct, when properly interpreted, in $\mathcal{H}^-$. The first is widely known to be correct when $z \in \mathcal{C}, z$ not a pole of the right-hand side. The second is not, since the right-hand side does not provide the correct analytic continuation of (2.21).

3. Preliminary results. All of our work will involve cut operators operating on products of Gaussian hypergeometric functions. The basic operand is the product

$$(3.1)\qquad \begin{aligned} f_{k,l,n}(a,b,\delta;z) \equiv f(z) &= {}_2F_1\left(\begin{matrix}a,\delta\\\delta+b\end{matrix};z\right)\\ &\quad\times {}_2F_1\left(\begin{matrix}1-a-k,-\delta-n\\1-\delta-b-l-n\end{matrix};z\right), \qquad k,l \in \mathcal{Z},\quad n \in \mathcal{N}. \end{aligned}$$

Define

$$(3.2)\qquad C = (-1)^k(a)_k(b)_l.$$

THEOREM 3.1. *For $k \geq l \geq 0$ or for $0 \geq k \geq l$,*

$$(3.3)\qquad \begin{aligned} &(\delta+b)_{n+l}\Pi^z_{k,n+l}\{f_{k,l,n}(a,b,\delta;z)\}\\ &\quad = C(1+\delta-a)_{n-k}\Pi^z_{k,n+l}\{z^n f_{-l,-k,n}(1-b,1-a,\delta;1/z)\}. \end{aligned}$$

Proof. First, assume $k \geq l \geq 0$. Obviously, we can assume that $n+l \geq k$, otherwise the result is trivial. Denote the left-hand side of (3.3) by G, the right-hand side by H. Then for $\delta \in \mathcal{Z}, -n \leq \delta \leq 0, G = H$. (Just turn both series around, using (2.10).) We can write

$$(3.4)\qquad G = H + V(\delta)_{n+1}.$$

Also,

$$(3.5)\qquad H = C\Pi^z_{k,n+l}\left\{\sum_{m=0}^{\infty} z^{n-m}\sum_{j=0}^{m} C_{m,j}(\delta)_j(-\delta-n)_{m-j}(1+\delta+j-a)_{n-m-k}\right\},$$

where $C_{m,j}$ is independent of δ and z. H is a polynomial of degree at most $n-k$ in δ if $k \leq n$, and is identically zero if $k > n$. One can similarly show G is a polynomial of degree at most $n+l$. Thus V is a polynomial of degree at most $l-1$. The result is immediate if $l = 0$. Assume $l > 0$. Obviously

$$V = \Pi^{\delta}_{0,l-1}\left\{\Pi^{z}_{k,n+l}\left\{\frac{(\delta+b)_{n+l}}{(\delta)_{n+l}}(1-z)^{k-1}\sum_{m=0}^{\infty}A_m(\delta)\left(\frac{z}{z-1}\right)^m\right\}\right\}, \tag{3.6}$$
$$A_m(\delta) = O(\delta^{-m}).$$

We consider all δ series to be members of $\mathcal{H}^-$, all z series to be members of $\mathcal{H}^+$. In the previous step we have used Euler-type transformations on the ${}_2F_1$'s; see [Erd, vol. 1, p. 64, (22)]. It is easily shown that these transformations are valid in $\mathcal{H}^+$. Because of the order estimate on A_m we have

$$V = \Pi^{\delta}_{0,l-1}\left\{\Pi^{z}_{k,n+l}\left\{\frac{(\delta+b)_{n+l}}{(\delta)_{n+1}}\sum_{m=0}^{l-1}A_m(\delta)(z)^m(1-z)^{k-m-1}\right\}\right\}. \tag{3.7}$$

(This step requires commutativity of the cut operators.) Since $k \geq l$ the quantity in the inner brackets is a polynomial in z of degree at most $k-1$, and consequently the above expression is zero. The first statement of the theorem is established.

The second statement follows from the first. We let $z \to 1/z$ and use property (2.3) of the cut operator, then property (2.4) with $p = l-k$. Finally, we replace k by $-l, l$ by $-k, a$ by $1-b, b$ by $1-a$ and interchange the two sides of the equation. □

The above result will produce an abundance of ${}_4F_3(1)$ identities when coefficients of like powers of z^r on both sides are equated. We shall display these identities later. However, there is one case not covered by the above result, that is, when the coefficient of z^n is selected in the case $k = n+1$. The resulting ${}_4F_3(1)$ identity is a generalization of one that occurs in the theory of the associated Jacobi polynomials; see [Wim1]. The proof involves a delicate and interesting umbral calculus argument.

THEOREM 3.2. *For* $0 \leq l \leq n$,

$$\Pi^{z}_{n,n}\{f_{n+1,l,n}(a,b,\delta;z)\} = \frac{(-z)^n}{(\delta-a)(\delta+b)_{n+l}}[(\delta)_{n+1}(\delta+b-a)_l - (a)_{n+1}(b)_l]. \tag{3.8}$$

Proof. Using an argument similar to that of Theorem 3.1, we can write

$$(\delta+b)_{n+l}\Pi^{z}_{n,n}\{f_{n+1,l,n}(a,b,\delta;z)\} = C\frac{z^n}{(\delta-a)} + \frac{V}{(\delta-a)}(\delta)_{n+1}, \tag{3.9}$$

where V is of degree at most l in δ. Again as before, we get

$$V = \Pi^{\delta}_{0,l}\left\{\frac{(\delta-a)(\delta+b)_{n+l}}{(\delta)_{n+1}}\Pi^{z}_{n,n}\left\{(1-z)^n\sum_{m=0}^{\infty}A_m(\delta)\left(\frac{z}{z-1}\right)^m\right\}\right\}, \tag{3.10}$$
$$A_m(\delta) = O(\delta^{-m}).$$

Because of the order estimate on A_m we can truncate the sum after $l+1$ terms. Since $l \leq n$, the quantity inside the inner brackets is then a polynomial in z of degree n or

less. By isomorphism, we may determine the coefficient of z^n by considering z to be a complex variable λ, dividing by λ^n and letting $\lambda \to \infty$.

(3.11)
$$\begin{aligned} V &= z^n \Pi_{0,l}^{\delta} \left\{ \frac{(\delta-a)(\delta+b)_{n+l}}{(\delta)_{n+1}} \lim_{\lambda\to\infty} \left\{ \frac{(1-\lambda)^n}{\lambda^n} \sum_{m=0}^{l} A_m(\delta) \left(\frac{\lambda}{\lambda-1}\right)^m \right\} \right\} \\ &= (-z)^n \Pi_{0,l}^{\delta} \left\{ \frac{(\delta-a)(\delta+b)_{n+l}}{(\delta)_{n+1}} \left\{ \sum_{m=0}^{l} A_m(\delta) \right\} \right\} \\ &= (-z)^n \Pi_{0,l}^{\delta} \left\{ \frac{(\delta-a)(\delta+b)_{n+l}}{(\delta)_{n+1}} \left\{ \sum_{m=0}^{\infty} A_m(\delta) \right\} \right\} \\ &= (-z)^n \Pi_{0,l}^{\delta} \left\{ \frac{(\delta-a)(\delta+b)_{n+l}}{(\delta)_{n+1}} {}_2F_1\left(\begin{matrix} a, b \\ \delta+b \end{matrix} ; 1 \right) \times {}_2F_1\left(\begin{matrix} -n-a, 1-b-l \\ 1-\delta-b-l-n \end{matrix} ; 1 \right) \right\}. \end{aligned}$$

Using Theorem 2.1 we represent the ${}_2F_1$'s above by the umbral gamma functions of definition (2.11). We find a substantial cancellation of exponential factors and conclude that

$$V = (-z)^n \Pi_{0,l}^{\delta} \{ (\delta+b-a)_l \} = (-z)^n (\delta+b-a)_l. \tag{3.12}$$

Putting this back into (3.10) gives the theorem. □

Note explicit polynomial expressions for the quantities $\Pi_{r,s}^{z}$ can be obtained by using the formula (2.8) for multiplying hypergeometric series.

4. Some hypergeometric identities. Multiplying out the relevant hypergeometric functions using formula (2.8) of §1 gives

$$\begin{aligned} &z^{-r} \Pi_{r,r}^{z} \{ f_{k,l,n}(a,b,\delta;z) \} \\ &\quad = \frac{(1-a-k)_r (-\delta-n)_r}{r!(1-\delta-b-l-n)_r} {}_4F_3\left(\begin{matrix} -r, a, \delta, \delta+b+l+n-r \\ \delta+b, a+k-r, \delta+n+1-r \end{matrix} ; 1 \right). \end{aligned} \tag{4.1}$$

Thus Theorem 3.1 yields

(4.2)
$$\begin{aligned} &{}_4F_3\left(\begin{matrix} -r, a, \delta, \delta+b+l+n-r \\ \delta+b, a+k-r, \delta+n+1-r \end{matrix} ; 1 \right) \\ &\quad = E \times {}_4F_3\left(\begin{matrix} r-n, 1-b, \delta, \delta+1-a-k+r \\ \delta+1-a, 1-b-l-n+r, \delta+1+r \end{matrix} ; 1 \right), \qquad 0 \le l \le k \le r \le n+l, \end{aligned}$$

where

$$E = \frac{(-1)^{k+r} r! (\delta+1)_{n-r} (b)_l (b+l)_{n-r} (\delta+1-a)_{r-k}}{(n-r)! (\delta+1)_r (a+k-r)_{r-k} (\delta+b)_{l+n-r}}. \tag{4.3}$$

(We interpret $E = 0$ when $r > n$.) Note the functions are not Saalschutzian, so this identity cannot be derived from known ${}_4F_3(1)$ identities, such as those given in [Bai, p. 56]. A referee has pointed out that this identity can be proved trivially. First,

assume δ is an integer and none of the other numerator or denominator parameters are integers. Reverse the summation index from j, say, to, $\delta - j$. Write the answer in terms of gamma functions and then let $r \to$ a positive integer. Finally free the parameter δ, which is permissible, since both sides are rational functions of δ.

Utilizing Theorem 3.2 gives the (known) identity

(4.4)
$$ {}_4F_3\left(\begin{matrix}-n, a, \delta, \delta+b+l \\ \delta+b, a+1, \delta+1\end{matrix}; 1\right) = \frac{n!a\delta}{(\delta-a)(\delta+b)_l}\left[\frac{(\delta+b-a)_l}{(a)_{n+1}} - \frac{(b)_l}{(\delta)_{n+1}}\right], \qquad 0 \le l \le n. $$

The reference [Wim1], shows the result also holds for $l = n+1$. However, it does *not* hold for $l > n+1$. (Just let $a \to \infty$. The result should be finite.) This is a strange identity. The left-hand side is a polynomial of degree n in l, and it would be tempting to assume the result holds for all l since it holds for $n+1$ consecutive values of l. However the right-hand side is not a polynomial in l. It is a transcendental function.

A referee has pointed out that this identity is actually a consequence of earlier work of one of the authors and J. Fields [Fie]. It has extensions to q-series, which are discussed in [Gas].

5. Two-point Padé approximants. Our main tool in this section will be Theorem 3.1.

Define

(5.1)
$$ \begin{aligned} P_{k,l,n}(\delta, z) &\equiv P_n(\delta, z) = \Pi^z_{0,n+l}\{f_{k,l,n}(a, b, \delta; z)\} \\ &= \Pi^z_{0,n+l}\left\{{}_2F_1\left(\begin{matrix}a, \delta \\ \delta+b\end{matrix}; z\right) \times {}_2F_1\left(\begin{matrix}1-a-k, -\delta-n \\ 1-\delta-b-l-n\end{matrix}; z\right)\right\}. \end{aligned} $$

Formula (3.3) shows that P_n is of *exact degree* n if $0 \le k \le n$. Since $\Pi_{0,n+l} = \Pi_{0,k-1} + \Pi_{k,n+l}$, the same formula shows that it is of degree *at most* n if $k = n+1$. Thus

(5.2)
$$ P_n(\delta, z) = \Pi^z_{0,n}\{f_{k,l,n}(a, b, \delta; z)\}, \qquad 0 \le l \le k \le n+1. $$

THEOREM 5.1. *For $0 \le l \le k \le n$ we have the two-point Padé approximant,*

(5.3)
$$ \frac{zP_{n-1}(\delta+1, z)}{P_n(\delta, z)} = \begin{cases} \dfrac{z\,{}_2F_1\left(\begin{matrix}a, \delta+1 \\ \delta+b+1\end{matrix}; z\right)}{{}_2F_1\left(\begin{matrix}a, \delta \\ \delta+b\end{matrix}; z\right)} + O(z^{n+l+1}), & z \to 0, \\ \dfrac{(\delta+b){}_2F_1\left(\begin{matrix}1-b, \delta+1 \\ \delta+2-a\end{matrix}; \frac{1}{z}\right)}{(1+\delta-a){}_2F_1\left(\begin{matrix}1-b, \delta \\ \delta+1-a\end{matrix}; \frac{1}{z}\right)} + O(z^{-n+k}), & z \to \infty. \end{cases} $$

Proof. We have

(5.4)
$$ \begin{aligned} P_n(\delta, z) &= \Pi^z_{0,n+l}\{f_{k,l,n}(a, b, \delta; z)\} \\ &= {}_2F_1\left(\begin{matrix}a, \delta \\ \delta+b\end{matrix}; z\right) \times {}_2F_1\left(\begin{matrix}1-a-k, -\delta-n \\ 1-\delta-b-l-n\end{matrix}; z\right) + O(z^{n+l+1}), \qquad z \to 0. \end{aligned} $$

Replacing n by $n-1, \delta$ by $\delta+1$, and dividing this equation by the original produces the first part of the theorem.

Next, we have

$$
\begin{aligned}
z^n P_n(\delta, 1/z) &= z^n[\Pi^z_{0,n+l}\{f_{k,l,n}(a,b,\delta;z)\}]|_{z\to 1/z} \\
&= z^n[\Pi^z_{0,k-1}\{f_{k,l,n}(a,b,\delta;z)\}]|_{z\to 1/z} \\
&\quad + z^n[\Pi^z_{k,n+l}\{f_{k,l,n}(a,b,\delta;z)\}]|_{z\to 1/z} \\
&= \mathrm{O}(z^{n-k+1}) + \frac{C(1+\delta-a)_{n-k}}{(\delta+b)_{n+l}} z^n \\
&\quad \times [\Pi^z_{k,n+l}\{z^n f_{-l,-k,n}(1-b,1-a,\delta;1/z)\}]|_{z\to 1/z} \\
&= \mathrm{O}(z^{n-k+1}) + \frac{C(1+\delta-a)_{n-k}}{(\delta+b)_{n+l}} {}_2F_1\left(\begin{matrix}1-b,\delta\\ \delta+1-a\end{matrix};z\right) \\
&\quad \times {}_2F_1\left(\begin{matrix}b+l,-\delta-n\\ a+k-\delta-n\end{matrix};z\right), \quad z\to 0.
\end{aligned} \tag{5.5}
$$

For the last step we used properties (2.3), (2.4). Replacing z by $1/z$ and using the same n, δ argument as above, we get the second part of the theorem. □

Put $k=l=0$ above to get Hendriksen's result; $k=l=n$ gives the $[n-1,n]$ Padé approximant given by Wimp. Letting $z\to 1/z$ and redefining parameters will yield results valid for $-n\le l\le k\le 0$, hence, the $[n-1,n]$ Padé approximant to the second ratio in (5.3) about $z=\infty$. It is interesting that the hypergeometric functions in the numerator, respectively, the denominator, of the first line of the theorem are solutions about ∞ of the same hypergeometric equation as the functions in the numerator, respectively, the denominator, of line one, which are solutions about 0.

Theorem 5.1 is a little unsatisfying because it does not yield an $[n,n]$ Padé approximant about $z=0$ when $l=n$. However, we may obtain such a result, in fact, a *family* of one-point approximants by using a trick. The symmetry relation

$$
f_{k+1,l,n}(a,b,\delta;z) = f_{n+1,n+l-k,k}(\delta, b+\delta-a; a; z) \tag{5.6}
$$

is easily verified. Define

$$
\begin{aligned}
W_{k,l,n}(\delta,z) &= \Pi^z_{0,n+l}\{f_{k,l,n}(a,b,\delta;z)\} \\
&= \Pi^z_{0,n+l}\left\{{}_2F_1\left(\begin{matrix}a,\delta\\ \delta+b\end{matrix};z\right) \times {}_2F_1\left(\begin{matrix}1-a-k,-\delta-n\\ 1-\delta-b-l-n\end{matrix};z\right)\right\}.
\end{aligned} \tag{5.7}
$$

We have

$$
\begin{aligned}
W_{n+1,l,n-1}(\delta,z) &= \Pi^z_{0,n+l-1}\{f_{n+1,l,n-1}(a,b,\delta;z)\} \\
&= \Pi^z_{0,n+l-1}\{f_{n,l-1,n}(\delta,a,b+\delta-a;z)\} \\
&= \Pi^z_{0,n-1}\{f_{n,l-1,n}(\delta,a,b+\delta-a,z)\} \\
&\quad + \Pi^z_{n,n+l-1}\{f_{n,l-1,n}(\delta,a,b+\delta-a,z)\},
\end{aligned} \tag{5.8}
$$

and Theorem 3.1 tells us that this is a polynomial in z of degree n at most for $1\le l\le n+1$. Proceeding as in the proof of the previous theorem, we have the following result.

THEOREM 5.2.

$$\frac{W_{n+1,l,n-1}(\delta+1,z)}{W_{n+1,l,n}(\delta,z)} = \frac{{}_2F_1\left(\begin{matrix}a,\delta+1\\ \delta+b+1\end{matrix};z\right)}{{}_2F_1\left(\begin{matrix}a,\delta\\ \delta+b\end{matrix};z\right)} + O(z^{n+l}), \qquad 1 \le l \le n+1. \tag{5.9}$$

When $l = n+1$, both numerator and denominator are of exact degree n, and the result is the desired $[n,n]$ Padé approximant.

A number of interesting cases of Theorem 5.1 follow by taking limits. We give several of these. First, let $z \to z/a, a \to \infty$. Put $l = k$ (no generality is obtained by doing otherwise) and define $c = b + \delta$. Next use the formulas in [Erd, vol. 1, 6.3, 6.5, 6.6] liberally. Theorems 5.1 and 5.2 may be combined in this case to give the following corollary.

COROLLARY 5.1. *Let*

$$Q_n(c,z) = \Pi^z_{0,n+k}\{\Phi(b,c;-z) \times \Phi(1-b-k, 1-c-k-n; z)\}. \tag{5.10}$$

Then for $0 \le k \le n+1$,

$$\frac{Q_{n-1}(c+1,z)}{Q_n(c,z)} = \begin{cases} \dfrac{\Phi(c+1-b,c+1;z)}{\Phi(c-b,c;z)} + O(z^{n+k}), & z \to 0, \\ \dfrac{-c\Psi(c+1-b,c+1;z)}{\Psi(c-b,c;z)} + O(z^{-n+k-1}), & z \to \infty. \end{cases} \tag{5.11}$$

Remarks. (i) The Φ and Ψ functions are the usual confluent hypergeometric functions. The ratio of Ψ functions must be interpreted as a umbral series whose coefficients are given formally by taking the ratio of the asymptotic series representing the Ψ functions.

(ii) For $k = n+1$, the numerator on the left of (5.11) is of degree n, hence this case yields the $[n,n]$ Padé approximant; $k = n$ yields the $[n-1,n]$ Padé approximant.

(iii) One may instead let $z \to zb$ in Theorems 5.1 and 5.2 and let $b \to \infty$. This amounts to interchanging Φ and Ψ, z and $1/z$. One then gets the $[n,n]$ and $[n-1,n]$ Padé elements for Ψ as $z \to \infty$.

In this result let $b = \nu + 1/2, c = 2\nu + 1, z \to 2z$. Then apply the formulas in [Erd, vol. 1, 6.9.1; vol. 2, p. 79, (19), (21)].

COROLLARY 5.2. *Let*

$$U_n(z) = \Pi^z_{0,n+k-1}\{\Phi(\nu+1/2, 2\nu+2; -2z) \times \Phi(1/2-\nu-k, -2\nu-k-n; 2z)\}, \tag{5.12}$$

$$V_n(z) = \Pi^z_{0,n+k}\{\Phi(\nu+1/2, 2\nu+1; -2z) \times \Phi(1/2-\nu-k, -2\nu-k-n; 2z)\}. \tag{5.13}$$

Then for $0 \le k \le n+1$,

$$\frac{U_n(z)}{V_n(z)} = \begin{cases} 1 + \dfrac{\mathrm{I}_{\nu+1}(z)}{\mathrm{I}_\nu(z)} + O(z^{n+k}), & z \to 0, \\ 1 - \dfrac{\mathrm{K}_{\nu+1}(z)}{\mathrm{K}_\nu(z)} + O(z^{-n+k-1}), & z \to \infty. \end{cases} \tag{5.14}$$

Finally, in Corollary 5.1 put $z \to z/(1-b)$ and let $b \to \infty$. Then use the formulas in [Erd, vol. 1, p. 266, (18); p. 185, (2)]. The limit of the second member of (5.11) does not exist in $\mathcal{H}^-$, but the limit of the first exists in $\mathcal{H}^+$. We obtain an explicit formula for the diagonal and off-diagonal Padé approximants to a ratio of Bessel functions. (We believe this result to be new.)

COROLLARY 5.3. *Let*

$$H_n(c,z) = \sum_{m=0}^{\left[\frac{n+k}{2}\right]} \frac{(m-n-k)_m z^m}{m!(c)_m(1-c-n-k)_m}. \tag{5.15}$$

Then for $k = n, n+1$,

$$\frac{H_{n-1}(c+1,z)}{H_n(c,z)} = \frac{c\mathrm{I}_c(2\sqrt{z})}{\sqrt{z}\mathrm{I}_{c-1}(2\sqrt{z})} + O(z^{n+k}), \qquad z \to 0. \tag{5.16}$$

Here $[\cdot]$ *denotes the greatest integer function.*

Acknowledgments. We thank the referees for their suggestions.

REFERENCES

[Bai] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, 1935.

[Bak] G. A. BAKER, JR., AND P. GRAVES-MORRIS, *Padé approximations*, in Encyclopedia of Mathematics and Its Applications, Vol. 13, 14, Addison-Wesley, Reading, MA, 1981.

[Coo] S. C. COOPER, A. MAGNUS, AND J. H. MCCABE, *On the normal two-point Padé table*, J. Comp. Appl. Math., 16 (1986), pp. 371–380.

[Erd] A. ERDÉLYI, et al. Higher Transcendental Functions, McGraw-Hill, New York, Vol. 1–3, 1953.

[Fie] J. L. FIELDS AND J. WIMP, *Expansion of hypergeometric functions in hypergeometric functions*, Math. Comp., 15 (1961), pp. 390–395.

[Gar] A. M. GARCIA AND S. A. JONI, *A new expression for umbral operators and power series inversion*, Proc. Amer. Math. Soc., 64 (1977), pp. 179–185.

[Gas] G. GASPER AND R. RAHMAN, *Basic hypergeometric series*, Encyclopedia of Mathematics and Its Applications, Vol. 35, Addison-Wesley, Reading, MA, 1980.

[Gil] J. GILEWICZ, *Approximants de Padé*, Lecture Notes in Math., Vol. 667, Springer-Verlag, New York, 1978.

[Hen1] E. HENDRIKSEN AND H. VAN ROSSUM, *Orthogonal Laurent polynomials*, Proc. Kon. Ned. Acad. v. Wet., A 89(1); Indag. Mat., 48 (1986), pp. 17–36.

[Hen2] E. HENDRIKSEN, *A Weight Function for the Associated Jacobi Laurent Polynomials*, preprint, 1989.

[Hen3] ———, *Associated Jacobi Laurent polynomials*, preprint, 1989.

[Jon1] WILLIAM B. JONES AND W. J. THRON, *Continued fractions: analytic theory and applications*, Encyclopedia of Mathematics and Its Applications, Vol. 11, Addison-Wesley, Reading, MA, 1980.

[Jon2] ———, *Continued fractions in numerical analysis*, in Continued Fractions and Padé Approximants, C. Brezinski, ed., North-Holland, Amsterdam, 1990.

[Kho] A. N. KHOVANSKII, *The application of continued fractions and their generalizations to problems in approximation theory*, Noordhoff, Groningen, 1963.

[McC1] J. H. MCCABE AND J. A. MURPHY, *Continued fractions which correspond to power series expansions at two points*, J. Math. Inst. Appl., 17 (1976), pp. 233–247.

[McC2] J. H. MCCABE, *Two point Padé approximants and the quotient difference algorithm*, J. Comp. Appl. Math., 7 (1981), pp. 151–153.

[Rom1] S. ROMAN, *The algebra of formal series*, Adv. Math., 31 (1979), pp. 309–329.

[Rom2] S. ROMAN, *The algebra of formal series*, II, J. Math. Anal. Appl., 74 (1980), pp. 120–143.

[Rom3] S. ROMAN AND G. C. ROTA, *The umbral calculus*, Adv. Math., 27 (1978), pp. 95–188.

[Sla] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.

[Wal] H. S. WALL, *Analytic Theory of Continued Fractions*, Chelsea, New York, (1967).

[Wim1] J. WIMP, *Explicit formulas for the associated Jacobi polynomials and some applications*, Canadian J. Math., 39 (1987), pp. 983–1000.

[Wim2] ———, *Fields of formal Laurent series and applications to some problems in numerical analysis*, in Numerical and Applied Mathematics, C. Brezinski, ed., J. C. Baltzer Publishing, 1989.

[Wim3] J. WIMP AND R. BOYER, *Formal series and an algorithm for computing some special determinants with elements in a ring*, Appl. Numer. Math., 5 (1989), pp. 1–10.

[Wim4] J. WIMP, R. KLINE, A. GALARDI, AND D. COLTON, *Some preliminary observations on an algorithm for the computation of moment integrals*, J. Comp. Appl. Math., 19 (1987), 117–124.

[Wim5] J. WIMP, R. KLINE, AND A. GALARDI, *Algorithms based on difference equations of infinite order and the computation of Laplace-type integrals*, Computing, 37 (1986), pp. 1–18.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 775–790, July 1995

ELLIPTIC-PARABOLIC EQUATIONS WITH HYSTERESIS BOUNDARY CONDITIONS*

ULRICH HORNUNG† AND R. E. SHOWALTER‡

Abstract. A general porous-medium equation is uniquely solved subject to a pair of boundary conditions for the trace of the solution and a second function on the boundary. The use of maximal monotone graphs for the three nonlinearities permits not only the inclusion of the usual boundary conditions of Dirichlet, Neumann, or Robin type, including variational inequality constraints of Signorini type, but also dynamic boundary conditions and those that model hysteresis phenomena. It is shown that the dynamic is determined by a contraction semigroup in a product of L^1 spaces. Several examples and numerical results are described.

Key words. existence, uniqueness, porous-media equation, hysteresis, nonlinear boundary condition, semigroup

AMS subject classifications. 35K55, 35K65

1. Introduction.

We shall consider a degenerate-parabolic initial boundary value problem in the form

$$\frac{\partial}{\partial t}a(u) - \Delta u \ni f\ , \qquad x \in \Omega\ , \tag{1.1.a}$$

$$\frac{\partial}{\partial t}b(v) + \frac{\partial u}{\partial \nu} \ni g, \quad \text{and} \tag{1.1.b}$$

$$\frac{\partial u}{\partial \nu} \in c(v-u)\ , \qquad s \in \Gamma \tag{1.1.c}$$

for each $t > 0$ with initial values specified at $t = 0$ for $a(u)$ and $b(v)$. At each $t > 0$, u is a function on the bounded domain Ω in $\mathbb{R}^n$ with smooth boundary Γ, and v is a function on Γ. Each $a(\cdot), b(\cdot), c(\cdot)$ is a maximal monotone graph in $\mathbb{R} \times \mathbb{R}$ [7]. Our interest in (1.1) arises primarily from the fact that (1.1.b) together with (1.1.c) can represent *hysteresis* phenomena on the boundary. Specifically, consider the maximal monotone graph given by $\operatorname{sgn}(y) = \{-1\}$ for $y < 0$, $\operatorname{sgn}(0) = [-1, 1]$, and $\operatorname{sgn}(y) = \{1\}$ for $y > 0$. If we choose $c = \operatorname{sgn}^{-1}$, the *inverse* graph obtained by reflection of the coordinates, then (1.1.b) is an ordinary differential equation for $b(v)$ subject to the *constraint* (1.1.c),

$$u - 1 \le v \le u + 1\ .$$

If $g \equiv 0$, then the selection $w \in b(v)$, which realizes the equation (1.1.b), is constant except at the constraint; there the *control* $\frac{\partial u}{\partial \nu}$ forces the corresponding equality. Thus the relationship between u and $w \in b(v)$ is an example of a *generalized play* [14]. Furthermore, if we let $b = \operatorname{sgn}^+ \equiv \frac{1}{2}(1 + \operatorname{sgn})$, then (1.1.b) models a *perfect relay* [14]. Thus the system (1.1) consists of a generalized porous-media equation in the interior of Ω subject to a nonlinear dynamic Neumann constraint, which can contain

* Received by the editors March 26, 1992; accepted for publication (in revised form) December 15, 1993.

† Universität der Bundeswehr, D-85577 Neubiberg, Germany (ulrich@informatik.unibw-muenchen.de).

‡ Department of Mathematics, University of Texas at Austin, Austin, Texas 78712-1082 (show@math.utexas.edu). The research of this author was supported by a grant from the National Science Foundation.

hysteresis phenomena on the boundary. Here, w is the *internal state* of the hysteron, $v-u$ is the *order parameter*, and u is the external *input*. See [19] and [17] for further discussion of these terms and general perspectives on hysteresis.

Although the hysteresis effects obtained from the pair of graphs $b(\cdot)$, $c(\cdot)$ were our primary motivation, we were able to include the third graph $a(\cdot)$ with no essential additional difficulty. This is merely a reflection of the power of the method that was developed in [22]; this method permits the addition of gradient nonlinearities of p-Laplacian type in (1.1.a) as well as corresponding elliptic Laplace–Beltrami operators in (1.1.b) for the manifold Γ. See [18] for a treatment of the degenerate case $a(\cdot)=0$ corresponding to a Stefan problem on the boundary Γ. Adsorption in porous media may be governed by conditions on the surfaces of the solid material that are of hysteresis type. In that case, u is the concentration of a chemical species that is dissolved in the fluid occupying the pores, and w is its concentration on the surfaces once it has been adsorbed. If one assumes that the process is governed by certain thresholds, the adsorption rate shows a hysteresis phenomenon of the kind discussed in this paper. In [11] this idea is applied to homogenization of reactive transport through porous media. Additional papers that deal with problems closely related to those of the present paper are [2], [13], [24], [25], [26], [15], and [16], where parabolic problems with a hysteresis source term are studied.

A rather remarkable variety of boundary conditions is obtained in (1.1). For example, if $b\equiv 0$ we have an explicit *Neumann boundary condition*, and if $c\equiv 0$ it is homogeneous. (Clearly, any general solvability results cannot simultaneously allow $c=b=0$, because this forces $g=0$.) If $b(0)=\mathbb{R}$ (i.e., $b^{-1}=0$), then $v\equiv 0$ and we have a nonlinear Neumann constraint, and if $c(0)=\mathbb{R}$, we get $v=u$ on Γ and this satisfies a nonlinear *dynamic boundary condition* of Neumann type. If $b(0)=c(0)=\mathbb{R}$, we have the homogeneous *Dirichlet boundary condition*. For previous work on some of these various classes, we refer to [3], [4], [5], [6], [8], [20], and [23].

Our objective is to show that the dynamic of problem (1.1) is determined by a nonlinear semigroup of contractions on the Banach space $L^1(\Omega)\times L^1(\Gamma)$. The (negative of the) generator of this contraction semigroup is (the closure of) an operator $\mathbb{C}$ for which the *resolvent equation* $(I+\varepsilon\mathbb{C})([a,b])\ni[f,g]$ with $\varepsilon>0$ takes the form

$$a(u)-\varepsilon\Delta u\ni f\,,\qquad x\in\Omega\,, \tag{1.2.a}$$

$$b(v)+\varepsilon\frac{\partial u}{\partial\nu}\ni g\,,\quad\text{and} \tag{1.2.b}$$

$$\frac{\partial u}{\partial\nu}\in c(v-u)\,,\qquad s\in\Gamma \tag{1.2.c}$$

in the state space $L^1(\Omega)\times L^1(\Gamma)$. In order to motivate the essential estimates that are needed, consider the (much simpler) case of *functions* $a(\cdot)$, $b(\cdot)$, $c(\cdot)$. Multiply the respective equations by appropriate functions φ on Ω and ψ on Γ, and integrate to obtain

$$\int_\Omega\big(a(u)\varphi+\varepsilon\vec{\nabla}u\cdot\vec{\nabla}\varphi\big)\,dx+\int_\Gamma\big(b(v)\psi+\varepsilon c(v-u)(\psi-\varphi)\big)\,ds=\int_\Omega f\varphi\,dx+\int_\Gamma g\psi\,ds\,. \tag{1.3}$$

This leads to the variational formulation of (1.2) and a priori estimates. For example, if we choose $\varphi=\operatorname{sgn}(u)$, $\psi=\operatorname{sgn}(v)$ and can simultaneously obtain $\varphi=\operatorname{sgn}(a(u))$, $\psi=\operatorname{sgn}(b(v))$, then we (formally) obtain the stability estimate

$$\|a(u)\|_{L^1(\Omega)}+\|b(v)\|_{L^1(\Gamma)}\le\|f\|_{L^1(\Omega)}+\|g\|_{L^1(\Gamma)}\,. \tag{1.4}$$

For the special case $a(u) = u$, $b(v) = v$, we could choose $\varphi = u$, $\psi = v$ and obtain corresponding L^2-estimates. For this special case we shall show that the corresponding evolution is *parabolic* in $L^2(\Omega) \times L^2(\Gamma)$; the same holds for its *additive perturbation* (see (5.1)). For the general case, estimate (1.4) suggests that the *resolvent* $[f, g] \mapsto [u, v] \to [a(u), b(v)]$ is a contraction. Of course we must obtain such estimates on *differences* of solutions.

Our plan is the following: In §2 we formulate the boundary value problem (1.2) as a variational problem in Sobolev space and give sufficient conditions for which it is well posed. In §3 we show that (1.1) is governed by a contraction semigroup on $L^1(\Omega) \times L^1(\Gamma)$ by constructing the operator $\mathbb{C}$, as suggested by our formal calculation above. Section 4 consists of some numerical examples which illustrate the hysteresis phenomena. Additional examples appear in [12]. Finally, we note in §5 that a corresponding *additive* perturbation of independent interest corresponds to a subgradient in Hilbert space from which one obtains parabolic regularizing effects.

2. The resolvent problem. Our objective is to make the boundary value problem (1.2) precise and give sufficient conditions for it to be well posed. Let Ω be a bounded domain in $\mathbb{R}^n$ with smooth boundary $\Gamma = \partial\Omega$. Denote by $L^p(\Omega)$ the usual space of Lebesgue pth-power integrable (equivalence classes of) functions on Ω when $1 \le p < \infty$, and denote by $L^\infty(\Omega)$ the essentially bounded measurable functions. Let $C_0^\infty(\Omega)$ be the infinitely differentiable functions with compact support in Ω, let $H^m(\Omega)$ be the Hilbert space of functions in $L^2(\Omega)$ for which each partial derivative up to order m belongs to $L^2(\Omega)$, and denote by $H_0^m(\Omega)$ the closure in $H^m(\Omega)$ of $C_0^\infty(\Omega)$. See [1] for information on these *Sobolev spaces*. Specifically, the *trace* map γ which assigns boundary values is well defined, continuous, and linear from $H^1(\Omega)$ into $L^2(\Gamma)$ with dense range $B = H^{1/2}(\Gamma)$.

We consider the Laplacian as an elliptic differential operator in divergence form from $H_0^1(\Omega)$ to its *dual* $H^{-1}(\Omega)$. Thus, assume we are given $a_{ij} \in L^\infty(\Omega)$, $1 \le i, j \le n$, which are uniformly positive definite; there is a $c_0 > 0$ for which

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \ge c_0|\xi|^2 , \qquad \xi \in \mathbb{R}^n, \tag{2.1}$$

where $|\xi|^2 = \sum_{j=1}^n |\xi_j|^2$. Then $\mathcal{A} : H^1(\Omega) \to H^1(\Omega)'$ is defined by

$$\mathcal{A}u(\varphi) = \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}\partial_i u \partial_j \varphi \right) dx , \qquad u, \varphi \in H^1(\Omega) .$$

The *formal part* of $\mathcal{A}$ is its restriction to $C_0^\infty(\Omega)$, the distribution

$$Au \equiv \mathcal{A}u|_{C_0^\infty(\Omega)} = -\sum_{i,j=1}^{n} \partial_j(a_{ij}\partial_i u) \in H^{-1}(\Omega) . \tag{2.2}$$

The monotone graphs in system (1.1) will be given as subgradients of convex functions [10]. Thus assume each $\zeta_a, \zeta_b, \zeta_c$ is a convex lower-semicontinuous function from $\mathbb{R}$ into the nonnegative extended reals $\mathbb{R}_\infty^+ = [0, +\infty]$, $\zeta_a(0) = \zeta_b(0) = \zeta_c(0) = 0$. Throughout most of the following we shall assume that ζ_c is quadratically bounded:

$$\zeta_c(s) \le C(|s|^2 + 1) , \qquad s \in \mathbb{R}; \tag{2.3.c}$$

hence ζ_c is continuous on *all* of $\mathbb{R}$. By defining

$$Z_a(u) \equiv \int_\Omega \zeta_a\big(u(x)\big)\,dx\,, \qquad u \in L^2(\Omega)\,, \tag{2.4.a}$$

$$Z_b(v) \equiv \int_\Gamma \zeta_b\big(v(s)\big)\,ds\,, \qquad v \in L^2(\Gamma)\,, \tag{2.4.b}$$

we obtain a pair of proper, convex, lower-semicontinuous functions, $Z_a : L^2(\Omega) \to \mathbb{R}^+_\infty$ and $Z_b : L^2(\Gamma) \to \mathbb{R}^+_\infty$. (By "proper" we mean that a function has a finite value somewhere.) Also, we define such a function Z_c on the product space $H^1(\Omega) \times L^2(\Gamma)$ by

$$Z_c\big([u,v]\big) \equiv \int_\Gamma \zeta_c\big(v(s) - \gamma u(s)\big)\,ds\,, \qquad u \in H^1(\Omega)\,,\ v \in L^2(\Gamma)\,, \tag{2.4.c}$$

and Z_c is convex and continuous on $H^1(\Omega)\times L^2(\Gamma)$. The subgradients of these functions are easily computed by standard results [10]. Thus, we have $a \in \partial Z_a(u)$ in $L^2(\Omega)$ if and only if

$$a(x) \in \partial\zeta_a\big(u(x)\big) \qquad \text{a.e. } x \in \Omega\,, \tag{2.5.a}$$

and similarly we have $b \in \partial Z_b(v)$ in $L^2(\Gamma)$ exactly when

$$b(s) \in \partial\zeta_b\big(v(s)\big) \qquad \text{a.e. } s \in \Gamma. \tag{2.5.b}$$

Since imbedding $H^1(\Omega)$ into $L^2(\Omega)$ is continuous and dense and we identify $L^2(\Omega)$ with its dual, we have $L^2(\Omega) \subset H^1(\Omega)'$. Thus, $a \in \partial Z_a(u)$ in $L^2(\Omega)$ implies that the same holds in $H^1(\Omega)'$, but $a \in \partial Z_a(u)$ in $H^1(\Omega)'$ does *not necessarily* imply (2.5.a). We shall call a subgradient in $H^1(\Omega)'$ a *weak subgradient* and one in $L^2(\Omega)$ a *strong subgradient*. Finally, since $Z_c : H^1(\Omega) \times L^2(\Gamma) \to \mathbb{R}$ is a composition of continuous functions, we have from the *chain rule* [10] that its weak subgradient is characterized by $C \in \partial Z_c[u,v]$ in $H^1(\Omega)' \times L^2(\Gamma)$ if and only if $C = [-\gamma' c, c]$ with

$$c(s) \in \partial\zeta_c\big(v(s) - \gamma u(s)\big) \qquad \text{a.e. } s \in \Gamma\,. \tag{2.5.c}$$

The dual map γ' of $L^2(\Gamma)$ into $H^1(\Omega)'$ is given by

$$\gamma' g(\psi) = \int_\Gamma g \cdot \gamma\psi\,ds\,, \qquad g \in L^2(\Gamma)\,,\ \psi \in H^1(\Omega)\,.$$

The boundary value problem (1.2) can now be realized as a subgradient equation. To this end, set

$$Z[u,v] \equiv Z_a(u) + Z_b(v) + \tfrac{1}{2}\mathcal{A}u(u) + Z_c[u,v]\,, \qquad u \in H^1(\Omega)\,,\ v \in L^2(\Gamma). \tag{2.6}$$

Clearly, there is no loss of generality in taking $\varepsilon = 1$, so we shall do so for the remainder of this section. Then, Z is the sum of convex and lower-semicontinuous functions, Z is proper, the first two terms are independent, and the remaining two are continuous and defined everywhere. Thus, we can compute the weak subgradient term by term. From this it follows that

$$\partial Z\big([u,v]\big) \ni [f,g] \ \text{ in } \ H^1(\Omega)' \times L^2(\Gamma), \tag{2.7}$$

whenever we have $u \in H^1(\Omega)$, $v \in L^2(\Gamma)$, and there exists $a \in L^2(\Omega)$, b, and $c \in L^2(\Gamma)$ satisfying (2.5) and

$$a + \mathcal{A}u - \gamma' c = f \quad \text{in } H^1(\Omega)', \tag{2.8.a}$$

$$b + c = g \quad \text{in } L^2(\Gamma). \tag{2.8.b}$$

That is, the weak subgradient (2.7) follows from (2.5) and (2.8). Moreover, (2.7) is equivalent to (2.5) and (2.8) if the first two terms are both *strong* subgradients. This will always be the case (by the chain rule) when we assume bounds of the form

$$\zeta_a(s) \le C(|s|^2 + 1)\,, \tag{2.3.a}$$

$$\zeta_b(s) \le C(|s|^2 + 1)\,, \qquad s \in \mathbb{R}\,. \tag{2.3.b}$$

In order to show that (2.8.a) is equivalent to a partial differential equation in Ω and a boundary condition on Γ, we develop an appropriate *Green formula* for the operator $\mathcal{A}$ [21]. Use the formal part (2.2) to define the domain

$$D \equiv \{u \in H^1(\Omega) : Au \in L^2(\Omega)\}\,.$$

Note that if Γ and the coefficients in A are smooth, then $D = H^2(\Omega)$. Recall that we denote the range of the trace γ by B and that B is dense and continuously imbedded in $L^2(\Gamma)$. Thus we obtain the identification $L^2(\Gamma) \subset B'$.

LEMMA 1. *There is a unique linear operator $\partial_A : D \to B'$ such that $\mathcal{A}u = Au + \gamma' \partial_A u$ for $u \in D$. That is, we have for each $u \in D$,*

$$\mathcal{A}u(\varphi) = (Au, \varphi)_{L^2(\Omega)} + \partial_A u(\gamma\varphi)\,, \qquad \varphi \in H^1(\Omega)\,. \tag{2.9}$$

Proof. Since γ is a strict homomorphism of $H^1(\Omega)$ onto B, its dual γ' is an isomorphism of B' onto the annihilator $H_0^1(\Omega)^\perp$ in $H^1(\Omega)'$ of $H_0^1(\Omega)$, the kernel of γ. Thus, for each $u \in D$, the difference $\mathcal{A}u - Au$ belongs to $H_0^1(\Omega)^\perp$, so it equals $\gamma'(\partial_A u)$ for a unique $\partial_A u \in B'$. □

The identity (2.9) is a generalization of the classical Green theorem. If Γ is sufficiently smooth and ν denotes the unit outward normal on Γ, and if $u \in H^2(\Omega)$ and $a_{ij} \in C^1(\overline{\Omega})$, $1 \le i, j \le n$, then

$$\int_\Omega \sum_{i,j=1}^n a_{ij} \partial_i u \partial_j \varphi \, dx = \int_\Omega Au\varphi \, dx + \int_\Gamma \frac{\partial u}{\partial \nu} \gamma\varphi \, ds\,, \qquad \varphi \in H^1(\Omega)\,,$$

where Au is given by (2.2) and the normal derivative is given by

$$\frac{\partial u}{\partial \nu} = \sum_{j=1}^n \Big(\sum_{i=1}^n a_{ij} \partial_i u \Big) \nu_j \in L^2(\Gamma)\,.$$

We can thus regard ∂_A as an extension of $\frac{\partial u}{\partial \nu}$ to a (possibly) wider class of functions in D.

Consider (2.8.a) and assume $f \in L^2(\Omega)$. Applying it to $C_0^\infty(\Omega)$ shows that

$$a + Au = f \quad \text{in } L^2(\Omega)\,. \tag{2.10.a}$$

Since from (2.10.a) it follows that $u \in D$, we may use (2.8.a) and (2.9) to get

$$\partial_A u = c \quad \text{in } L^2(\Gamma)\,. \tag{2.10.c}$$

Then (2.8.b) is equivalent to

$$b + \partial_A u = g \quad \text{in} \quad L^2(\Gamma) \,. \tag{2.10.b}$$

This shows that (2.8) is equivalent to (2.10), and we have shown that the strong subgradient identity (2.7) is satisfied by a solution of the resolvent problem (1.2), namely, (2.5) and (2.10).

The following result gives sufficient conditions for the resolvent problem to be solvable and equivalent to (2.7) in $L^2(\Omega) \times L^2(\Gamma)$.

THEOREM 1. *Let the domain Ω with boundary $\Gamma = \partial\Omega$, the coefficients $\{a_{ij}\}$ satisfying* (2.1), *and the convex lower-semicontinuous functions $\zeta_a, \zeta_b, \zeta_c$ from $\mathbb{R}$ into $\mathbb{R}^+_\infty$ with $\zeta_a(0) = \zeta_b(0) = \zeta_c(0) = 0$ be given. Assume* (2.3.a)–(2.3.c) *and that for some $c_1 > 0$,* any two *of the following hold:*

$$\zeta_a(s) \geq c_1|s|^\alpha - C \quad \textit{with} \quad 1 < \alpha \leq 2 \,, \tag{2.11.a}$$

$$\zeta_b(s) \geq c_1|s|^2 - C \,, \qquad s \in \mathbb{R} \,, \tag{2.11.b}$$

$$\zeta_c(s) \geq c_1|s|^2 - C \,, \qquad s \in \mathbb{R} \,. \tag{2.11.c}$$

Then, for the proper, convex, and lower-semicontinuous $Z : H^1(\Omega) \times L^2(\Gamma) \to \mathbb{R}^+_\infty$ given by (2.4) *and* (2.6), *it follows that the subgradient ∂Z is surjective onto $H^1(\Omega)' \times L^2(\Gamma)$. Thus, for each triple $f \in L^2(\Omega)$, $g, h \in L^2(\Gamma)$, there exists a solution pair $u \in H^1(\Omega)$, $v \in L^2(\Gamma)$ and corresponding selections $a \in L^2(\Omega)$, $b, c \in L^2(\Gamma)$ satisfying* (2.5) *and*

$$a + Au = f \quad \textit{in} \quad L^2(\Omega), \tag{2.12.a}$$

$$b + c = g \quad \textit{in} \quad L^2(\Gamma), \tag{2.12.b}$$

$$\partial_A u - c = h \quad \textit{in} \quad L^2(\Gamma). \tag{2.12.c}$$

Proof. From Green's identity it follows that (2.12) is equivalent to

$$\begin{aligned} a + \mathcal{A}u - \gamma' c &= f + \gamma' h \quad \text{in} \quad H^1(\Omega)', \\ b + c &= g \qquad\qquad\quad \text{in} \quad L^2(\Gamma), \end{aligned}$$

and this, in turn, is equivalent to

$$\partial Z\big([u, v]\big) \ni [f + \gamma' h, g] \quad \text{in} \quad H^1(\Omega)' \times L^2(\Gamma).$$

These equivalences follow by the same calculations relating (2.7), (2.8), and (2.10). Thus it suffices to show that Z is *coercive* on $H^1(\Omega) \times L^2(\Gamma)$, i.e.,

$$\frac{Z([u, v])}{\|u\|_{H^1(\Omega)} + \|v\|_{L^2(\Gamma)}} \to \infty \quad \text{as} \quad \|u\|_{H^1(\Omega)} + \|v\|_{L^2(\Gamma)} \to \infty. \tag{2.13}$$

We shall verify (2.13). If the fraction in (2.13) is bounded, then we obtain for some constant K,

$$\int_\Omega \Big(\zeta_a(u) + \frac{c_0}{2}|\nabla u|^2\Big)\, dx + \int_\Gamma \big(\zeta_b(v) + \zeta_c(v - \gamma u)\big)\, ds \tag{2.14}$$

$$\leq K\big\{\|\nabla u\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)} + \|v\|_{L^2(\Gamma)}\big\} \,.$$

LEMMA 2. *There is a constant K_1 such that*

$$\|u\|_{L^2(\Omega)} \leq K_1\big(\|\nabla u\|_{L^2(\Omega)} + \|u\|_{L^1(\Omega)}\big) \,, \qquad u \in H^1(\Omega) \,. \tag{2.15}$$

Proof. Otherwise there is a sequence $\{u_n\}$ in $H^1(\Omega)$ for which $\|u_n\|_{L^2} = 1$, and the right side of (2.15) converges to zero. Then, $\{u_n\}$ is bounded in $H^1(\Omega)$, so (by passing to a subsequence) we have $u_n \rightharpoonup u$ in $H^1(\Omega)$ and $u_n \to u$ in $L^2(\Omega)$ by compactness. But then $u_n \rightharpoonup u$ in the weaker norm on the right of (2.15), so by uniqueness of weak limits, we have $u = 0$. Thus $u_n \to 0$ in $L^2(\Omega)$, a contradiction. □

Suppose we have the case of (2.11.a) and (2.11.b). Using Lemma 2, we replace $\|u\|_{L^2(\Omega)}$ by $\|u\|_{L^1(\Omega)}$ in (2.14), and then we have

$$c_1\|u\|^{\alpha}_{L^1(\Omega)} + \frac{c_0}{2}\|\nabla u\|^2_{L^2(\Omega)} + c_1\|v\|^2_{L^2(\Gamma)} \le K_2\{\|\nabla u\|_{L^2(\Omega)} + \|u\|_{L^1(\Omega)} + \|v\|_{L^2(\Gamma)} + 1\} .$$

From here it follows that $\|u\|_{H^1(\Omega)} + \|v\|_{L^2(\Gamma)}$ is bounded, so we have (2.13).

Suppose we have the case of (2.11.a) and (2.11.c). As before, we obtain

$$\begin{aligned} c_1\|u\|^{\alpha}_{L^1(\Omega)} + \frac{c_0}{2}\|\nabla u\|^2_{L^2(\Omega)} + c_1\|v - \gamma u\|^2_{L^2(\Gamma)} \\ \le K_2\{\|\nabla u\|_{L^2(\Omega)} + \|u\|_{L^1(\Omega)} + \|v - \gamma u\|_{L^2(\Gamma)} + \|\gamma u\|_{L^2(\Gamma)} + 1\} . \end{aligned}$$

Since γ is continuous from $H^1(\Omega)$ into $L^2(\Gamma)$, the term $\|\gamma u\|_{L^2(\Gamma)}$ can be absorbed in the first two terms by adjusting K_2, and then we are done.

Now consider the remaining case of (2.11.b) and (2.11.c); then, from (2.14) we have

$$\begin{aligned} \frac{c_0}{2}\|\nabla u\|^2_{L^2(\Omega)} + c_1\|v\|^2_{L^2(\Gamma)} + c_1\|v - \gamma u\|^2_{L^2(\Gamma)} \\ \le K_2\{\|\nabla u\|_{L^2(\Omega)} + \|\gamma u\|_{L^2(\Gamma)} + \|v\|_{L^2(\Gamma)} + 1\} , \end{aligned} \tag{2.16}$$

in which we have used either a Poincaré inequality or the argument of Lemma 2 to replace $\|u\|_{L^2(\Omega)}$ by $\|\gamma u\|_{L^2(\Gamma)}$. Using the inequality $2\alpha\beta \le \varepsilon\alpha^2 + \frac{1}{\varepsilon}\beta^2$ with $\alpha = \|v\|_{L^2(\Gamma)}$, $\beta = \|\gamma u\|_{L^2(\Gamma)}$, and $1 < \varepsilon < 2$, we obtain

$$\begin{aligned} (2-\varepsilon)\|v\|^2_{L^2(\Gamma)} + \Big(1 - \frac{1}{\varepsilon}\Big)\|\gamma u\|^2_{L^2(\Gamma)} &\le 2\|v\|^2_{L^2(\Gamma)} - 2(v, \gamma u)_{L^2(\Gamma)} + \|\gamma u\|^2_{L^2(\Gamma)} \\ &= \|v\|^2_{L^2(\Gamma)} + \|v - \gamma u\|^2_{L^2(\Gamma)} . \end{aligned}$$

Thus, we can replace $\|v - \gamma u\|^2_{L^2(\Gamma)}$ with $\|\gamma u\|^2_{L^2(\Gamma)}$ in (2.16) by adjusting c_1, so we obtain the coercivity condition (2.13) as before. □

3. The evolution problem. The goal in this section is to construct the generator of the nonlinear semigroup which corresponds to system (1.1). We shall assume the domain Ω in $\mathbb{R}^n$ with boundary $\Gamma = \partial\Omega$, the coefficients $\{a_{ij}\}$ in $L^\infty(\Omega)$, and the function $\zeta_c : \mathbb{R} \to \mathbb{R}^+_\infty$ are given as in §2.

Define a single-valued operator $\mathbb{C}_2$ on the Hilbert space $L^2(\Omega) \times L^2(\Gamma)$ as follows: $\mathbb{C}_2([u, v]) = [f, g]$ if and only if $[f, g] \in L^2(\Omega) \times L^2(\Gamma)$, and

$$u \in H^1(\Omega),\ \mathcal{A}u = f + \gamma' g \ \text{ in } \ H^1(\Omega)' , \ \text{ and} \tag{3.1.a}$$

$$v \in L^2(\Gamma),\ \partial\zeta_c(v - \gamma u) \ni g \ \text{ in } \ L^2(\Gamma) . \tag{3.1.b}$$

This is just (2.5) and (2.12) with $\zeta_a = \zeta_b = 0$ and $h = 0$, and it can be written as

$$\mathcal{A}u(\varphi) + \int_\Gamma c(\psi - \gamma\varphi)\, ds = \int_\Omega f\varphi\, dx + \int_\Gamma g\psi , \qquad \varphi \in H^1(\Omega) ,\ \psi \in L^2(\Gamma) ,$$

$$c \in \partial\zeta_c(v - \gamma u) \ \text{ in } \ L^2(\Gamma) .$$

According to Lemma 1, its value is given explicitly by $\mathbb{C}_2([u,v]) = [Au, \partial_A u]$. We shall first show that $\mathbb{C}_2$ is m-accretive and also a *subgradient* in $L^2(\Omega) \times L^2(\Gamma)$; this implies that the special case of the system (1.1) with $a = b =$ identity is well posed and *parabolic* (see §5). Then we shall show that the closure of the operator $\mathbb{C}$ which corresponds to system (1.2) with nonlinear $a(\cdot)$, $b(\cdot)$ is m-accretive in $L^1(\Omega) \times L^1(\Gamma)$.

PROPOSITION 1. *The function* $Z_2 : L^2(\Omega) \times L^2(\Gamma) \to \mathbb{R}^+_\infty$ *given by* (2.4.c) *and*

$$Z_2([u,v]) \equiv \frac{1}{2}\mathcal{A}u(u) + Z_c([u,v])$$

is proper convex and lower-semicontinuous. The (strong) subgradient is given by $\partial Z_2 = \mathbb{C}_2$.

Proof. The function Z_2 is clearly proper and convex. To see that it is lower-semicontinuous, note that if $[u_n, v_n] \to [u,v]$ in $L^2(\Omega) \times L^2(\Gamma)$ and $\{Z_2([u_n,v_n])\}$ is bounded, then $\{[u_n,v_n]\}$ is bounded in $H^1(\Omega) \times L^2(\Gamma)$, so for some subsequence, $\gamma u_n \to \gamma u$ (strongly) in $L^2(\Gamma)$ and Fatou's lemma yields the desired result. To compute the subgradient, use the termwise weak subdifferentiability to see that if $[f,g] \in \partial Z_2([u,v])$, then there exists a $c \in L^2(\Gamma)$ with (2.5.c) and

$$f(\varphi - u) + g(\psi - v) \le \mathcal{A}u(\varphi - u) + \int_\Gamma c\big(\psi - v - \gamma(\varphi - u)\big)\, ds\,, \qquad \varphi \in H^1(\Omega)\,,\ \psi \in L^2(\Gamma)\,;$$

this is easily seen to be equivalent to (3.1).

We develop additional estimates on $\mathbb{C}_2$ and begin with the following lemma.

LEMMA 3. *If* $\sigma : \mathbb{R} \to \mathbb{R}$ *is monotone and Lipschitz, and* $\sigma(0) = 0$, *then for each pair*

$$\mathbb{C}_2([u_j, v_j]) = [f_j, g_j]\,, \qquad j = 1, 2,$$

we have

$$\big(f_1 - f_2,\, \sigma(u_1 - u_2)\big)_{L^2(\Omega)} + \big(g_1 - g_2,\, \sigma(v_1 - v_2)\big)_{L^2(\Gamma)} \ge 0\,.$$

Proof. We use (3.1.a) to compute the above two terms. The composite $\sigma(u_1 - u_2)$ belongs to $H^1(\Omega)$ and by the chain rule we obtain

$$\mathcal{A}(u_1 - u_2)\big(\sigma(u_1 - u_2)\big) = \int_\Omega \sum_{i,j=1}^n a_{ij}\partial_i(u_1 - u_2)\partial_j(u_1 - u_2)\sigma'(u_1 - u_2)\, dx\,,$$

and this is nonnegative in view of (2.1) and the monotonicity of σ. Also, we have to check the remaining term

$$\int_\Gamma (c_1 - c_2)\big(\sigma(v_1 - v_2) - \sigma(\gamma u_1 - \gamma u_2)\big)\, ds,$$

but this is nonnegative because of (3.1.b) since $\partial\zeta_c$ is a monotone graph and σ is a monotone function. □

The special case of $\sigma(s) = s$ is just the observation that $\mathbb{C}_2$ is monotone in the Hilbert space $L^2(\Omega) \times L^2(\Gamma)$. Since $\mathbb{C}_2$ is single valued, we can permit σ to be *multivalued*.

PROPOSITION 2. *Let the domain* Ω *with boundary* Γ, *the coefficients* $\{a_{ij}\}$ *in* $L^\infty(\Omega)$ *satisfying* (2.1), *and the convex continuous function* $\zeta_c : \mathbb{R} \to \mathbb{R}^+_\infty$ *with* $\zeta_c(0) = 0$ *and* (2.3.c) *be given. Let* $j : \mathbb{R} \to \mathbb{R}^+_\infty$ *be convex and lower-semicontinuous, and let* $j(0) = 0$. *Then we have*

$$\big(\mathbb{C}_2[u_1, v_1] - \mathbb{C}_2[u_2, v_2], [\sigma_1, \sigma_2]\big)_{L^2(\Omega)\times L^2(\Gamma)} \ge 0 \tag{3.2}$$

for any *selections* $\sigma_1 \in \partial j(u_1 - u_2)$ *in* $L^2(\Omega)$ *and* $\sigma_2 \in \partial j(v_1 - v_2)$ *in* $L^2(\Gamma)$.

Proof. Consider the lower-semicontinuous convex function J on $L^2(\Omega) \times L^2(\Gamma)$ given by

$$J([u,v]) = \int_\Omega j(u(x))\,dx + \int_\Gamma j(v(s))\,ds\,, \qquad u \in L^2(\Omega)\,,\ v \in L^2(\Gamma)\,. \tag{3.3}$$

The subgradient of J is given by

$$\sigma \equiv [\sigma_1, \sigma_2] \in \partial J([u,v]) \ \text{ in } \ L^2(\Omega) \times L^2(\Gamma)$$

if and only if

$$\sigma([\varphi,\psi]) = \int_\Omega \sigma_1(x)\varphi(x)\,dx + \int_\Gamma \sigma_2(s)\psi(s)\,ds\,, \qquad \varphi \in L^2(\Omega)\,,\ \psi \in L^2(\Gamma)\,,$$

with $\sigma_1(x) \in \partial j(u(x))$ a.e. $x \in \Omega$, $\sigma_2(s) \in \partial j(v(s))$ a.e. $s \in \Gamma$. The Yoshida approximation J_ε of J is given by the same formula but with j replaced by j_ε. The derivative j_ε' is Lipschitz and monotone so Lemma 3 yields (3.2) in this special case. Thus, $\mathbb{C}_2$ is ∂J-monotone by Proposition 4.7 of [7] and the general case follows since the single-valued $\mathbb{C}_2$ is equal to its minimal section.

Remark. As a consequence of Proposition 2.17 of [7], we also obtain the following corollary.

COROLLARY 1. *Let* j *be given as above. Then* $\partial(J + Z_2) = \partial J + \partial Z_2$.

It follows that the special case of the boundary value problem (1.2) with $a = b = \partial j$ is well posed in $L^2(\Omega) \times L^2(\Gamma)$ when j satisfies an estimate of the form (2.11), because $J + Z_2$ is then coercive over $L^2(\Omega) \times L^2(\Gamma)$.

Next we construct the generator of the general system (1.1). This operator will be obtained by closing up the composition of $\mathbb{C}_2$ with the inverse of $[\partial\zeta_a, \partial\zeta_b]$ in $L^1(\Omega) \times L^1(\Gamma)$. As before, we shall always assume that (2.1) holds, $\zeta_a, \zeta_b, \zeta_c : \mathbb{R} \to \mathbb{R}^+_\infty$ are convex and lower-semicontinuous, and (2.3.c) holds.

DEFINITION. *The operator* $\mathbb{C}$ *in* $L^2(\Omega) \times L^2(\Gamma)$ *is defined as follows:* $\mathbb{C}([a,b]) \ni [f,g]$ *if there is a pair* $[u,v]$ *as in* (3.1) *and a pair* $a \in L^2(\Omega)$, $b \in L^2(\Gamma)$ *for which* $\mathbb{C}_2([u,v]) = [f,g]$ *and* $a \in \partial\zeta_a(u)$ *in* $L^2(\Omega)$, $b \in \partial\zeta_b(v)$ *in* $L^2(\Gamma)$.

Note that $Rg(I + \varepsilon\mathbb{C}) = L^2(\Omega) \times L^2(\Gamma)$ for $\varepsilon > 0$ in both the situation of Theorem 1 (i.e., (2.3.a) and (2.3.b)) and in the case of Corollary 1 with (2.11) and $\zeta_a = \zeta_b$.

LEMMA 4. *The operator* $\mathbb{C}$ *is accretive on* $L^1(\Omega) \times L^1(\Gamma)$.

Proof. Let $\varepsilon > 0$ and $(I + \varepsilon\mathbb{C})([a_j, b_j]) \ni [f_j, g_j]$ for $j = 1, 2$. Thus we have

$$\varepsilon\mathbb{C}_2([u_j, v_j]) = [f_j - a_j,\ g_j - b_j]\,, \qquad a_j \in \partial\zeta_a(u_j)\,,\ b_j \in \partial\zeta_b(v_j)$$

as above. We choose $j(s) = |s|$ so that $\partial j = \operatorname{sgn}$; then we use (3.3) with

$$\sigma_1 = \operatorname{sgn}_0(u_1 - u_2 + a_1 - a_2) \in \operatorname{sgn}(u_1 - u_2) \cap \operatorname{sgn}(a_1 - a_2),$$
$$\sigma_2 = \operatorname{sgn}_0(v_1 - v_2 + b_1 - b_2) \in \operatorname{sgn}(v_1 - v_2) \cap \operatorname{sgn}(b_1 - b_2)$$

to obtain

$$\|a_1 - a_2\|_{L^1(\Omega)} + \|b_1 - b_2\|_{L^1(\Gamma)} \le \|f_1 - f_2\|_{L^1(\Omega)} + \|g_1 - g_2\|_{L^1(\Gamma)}\,. \tag{3.4}$$

Of course the same procedure with the function $j(s) = s^+$ and its subgradient $\partial j = \operatorname{sgn}^+$ yields the *comparison estimate*

$$\|(a_1 - a_2)^+\|_{L^1(\Omega)} + \|(b_1 - b_2)^+\|_{L^1(\Gamma)} \le \|(f_1 - f_2)^+\|_{L^1(\Omega)} + \|(g_1 - g_2)^+\|_{L^1(\Gamma)}\,. \tag{3.5}$$

This leads to the following L^∞ estimates.

COROLLARY 2. *If* $(I+\varepsilon\mathbb{C})([a,b]) \ni [f,g]$ *and* $\|f^+\|_{L^\infty(\Omega)}+\|g^+\|_{L^\infty(\Gamma)} \in Rg(\partial\zeta_a+\partial\zeta_b)$, *then*

$$\|a^+\|_{L^\infty(\Omega)} \le \|f^+\|_{L^\infty(\Omega)}, \quad \|b^+\|_{L^\infty(\Gamma)} \le \|g^+\|_{L^\infty(\Gamma)}. \tag{3.6}$$

Proof. Set $a_2 = \|f^+\|_{L^\infty(\Omega)}$, $b_2 = \|g^+\|_{L^\infty(\Gamma)}$, and choose k such that $\partial\zeta_a(k) \ni a_2$ and $\partial\zeta_b(k) \ni b_2$. With $u(x) = k$, $v(s) = k$ in the definition of $\mathbb{C}$, we have $(I + \varepsilon\mathbb{C})([a_2,b_2]) \ni [a_2,b_2]$, so we can apply (3.5) to get $\|(a-a_2)^+\|_{L^1(\Omega)}+\|(b-b_2)^+\|_{L^1(\Gamma)} = 0$. □

The same result holds for the "negative parts," and by adding the corresponding estimates, we obtain estimate (3.5) with the "positive part" deleted throughout.

LEMMA 5. *Assume that any two parts of* (2.11) *hold. Then for any* $\varepsilon > 0$ *and* $[f,g] \in L^\infty(\Omega) \times L^\infty(\Gamma)$ *with* $\|f\|_{L^\infty(\Omega)} + \|g\|_{L^\infty(\Gamma)} \in Rg(\partial\zeta_\alpha + \partial\zeta_b)$, *there exists a unique* $[a,b]$ *such that* $(I + \varepsilon\mathbb{C})([a,b]) \ni [f,g]$ *and*

$$\|a\|_{L^\infty(\Omega)} \le \|f\|_{L^\infty(\Omega)}, \quad \|b\|_{L^\infty(\Gamma)} \le \|g\|_{L^\infty(\Gamma)}. \tag{3.7}$$

Proof. Modify ζ_a to replace $\partial\zeta_a$ by its *truncation*

$$\partial\zeta_a^m(s) = \begin{cases} \{\min\{r,m\} : r \in \partial\zeta_a(s)\} & \text{if } s \ge 0, \\ \{\max\{r,-m\} : r \in \partial\zeta_a(s)\} & \text{if } s < 0, \end{cases}$$

where $m = \max\{\|f\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\Gamma)}\}$. Thus $\partial\zeta_a^m$ has bounded range, so ζ_a^m satisfies (2.3.a). Likewise, modify ζ_b to obtain ζ_b^m satisfying (2.3.b). By Theorem 1, there is a unique solution $[a,b] \in L^2(\Omega) \times L^2(\Gamma)$ of $(I + \varepsilon\mathbb{C})([a,b]) \ni [f,g]$ with the modified functions ζ_a^m, ζ_b^m. This solution satisfies (3.7), so (2.5) holds since the modified functions agree with the original ones for these values of a and b. □

We summarize the above construction in the following.

THEOREM 2. *Assume we are given the domain* Ω *with boundary* Γ *as above, the coefficients* $\{a_{ij}\}$ *in* $L^\infty(\Omega)$ *satisfying* (2.1), *and the three convex, lower-semi-continuous functions* $\zeta_a, \zeta_b, \zeta_c$ *from* $\mathbb{R}$ *into* $\mathbb{R}^+_\infty$ *satisfying* $\zeta_a(0) = \zeta_b(0) = \zeta_c(0) = 0$ *and any two of* (2.11).

(a) *If either* (2.3.a)–(2.3.c) *holds or* $\zeta_a = \zeta_b$ *and* (2.3.c) *holds, then* $Rg(I + \varepsilon\mathbb{C}) = L^2(\Omega) \times L^2(\Gamma)$.

(b) *If* $Rg(\partial\zeta_a + \partial\zeta_b) = \mathbb{R}$, *then* $Rg(I + \varepsilon\mathbb{C}) \supset L^\infty(\Omega) \times L^\infty(\Gamma)$.

In both of these cases, the closure $\overline{\mathbb{C}}$ *of* $\mathbb{C}$ *in* $L^1(\Omega) \times L^1(\Gamma)$ *is m-accretive.*

Proof. Part (a) is implicit in Theorem 1 and Corollary 1. For part (b), we apply Lemma 5 and note that we have that $\|\partial_A u\|_{L^\infty(\Gamma)} \le \frac{2}{\varepsilon}\|g\|_{L^\infty(\Gamma)}$ from (3.7). Thus we may replace $\partial\zeta_c$ by its truncation $\partial\zeta_c^{2m/\varepsilon}$, and the corresponding convex $\zeta_c^{2m/\varepsilon}$ satisfies (2.3.c). □

Since $\overline{\mathbb{C}}$ is m-accretive, it follows from the Crandall–Liggett theorem [9] that the abstract Cauchy problem

$$\begin{aligned} &\tilde{a}'(t) + \overline{\mathbb{C}}\big(\tilde{a}(t)\big) \ni \tilde{f}(t)\,, \qquad 0 \le t \le T\,, \\ &\tilde{a}(0) = \tilde{a}_0 \end{aligned}$$

has an *integral solution* $\tilde{a}(t) = [a(t), b(t)]$ in $C([0,T], L^1(\Omega) \times L^1(\Gamma))$ which is unique; see also [3]. This solution can be obtained as the uniform limit of step functions obtained from the implicit difference scheme

$$[a^n, b^n] - [a^{n-1}, b^{n-1}] + h\,\overline{\mathbb{C}}\big([a^n, b^n]\big) \ni h[f^n, g^n]\,, \qquad 1 \le n \le N,$$

with step $h = T/N$ and $[a^0, b^0] = \tilde{a}_0 \in \mathrm{dom}(\overline{\mathbb{C}})$. This provides a generalized solution of the degenerate parabolic system

$$\frac{\partial a}{\partial t} + Au = f \ , \ a \in \partial\zeta_a(u) \ \text{ in } \ L^1(\Omega) \ ,$$
$$\frac{\partial b}{\partial t} + \frac{\partial u}{\partial \nu} = g \ , \ b \in \partial\zeta_b(v) \ , \ \frac{\partial u}{\partial \nu} \in \partial\zeta_c(v - \gamma u) \ \text{ in } \ L^1(\Gamma) \ ,$$

with initial data

$$a(x,0) = a^0(x) \quad \text{a.e.} \quad x \in \Omega,$$
$$b(s,0) = b^0(s) \quad \text{a.e.} \quad s \in \Gamma$$

as desired.

4. Examples. For the following numerical examples we have modified the initial boundary value problem (1) in that we assume the boundary Γ of the domain Ω is the union of two parts, namely, $\Gamma = \Gamma_D \cup \Gamma_H$. We prescribe Dirichlet data $u = u_D = h(t)$ on Γ_D and use the hysteresis boundary conditions (1.1.b), (1.1.c) on Γ_H. The modification of the theorems, such that this case is also covered, is obvious.

We consider a multiple of the signum function

$$b = \frac{1}{2}\,\mathrm{sgn}$$

($\varepsilon = 0$) or a smooth approximation thereof, namely,

$$b_\varepsilon(z) = \frac{1}{2}\frac{z}{\varepsilon + |z|},$$

and the inverse of the signum function

$$c(z) = \mathrm{sgn}^{-1}(z).$$

For the following examples we simplify by using $a(u) = u$ and $f, g = 0$. We are going to use the function

$$h(t) = \alpha 2^{-t/\beta} \sin(2\pi\omega t)$$

(with $\alpha, \beta, \omega > 0$). The initial values are all zero in the examples. As a numerical method, we have used the standard time-explicit difference scheme with constant step-sizes in x and t. Additional details and examples can be found in [12].

Example 1. As a one-dimensional example, let $\Omega = (0,1)$, $\Gamma_H = \{0\}$, $\Gamma_D = \{1\}$. We assume $u_D(t) = h(t)$ with $\alpha = 4$, $\beta = 10$, $\omega = 1/5$, and $\varepsilon = 0$. Figure 1 shows u and the selection $w \in b(u)$ at $x = 1$ as a function of time; the dotted line is the function h and w is the solid line bounded by $1/2$. Figure 2 shows w versus u; the oblique lines that cut the corners are a result of the discretization of time. This has the typical form of a *perfect relay*.

Example 2. The following is an example in two dimensions. We take $\Omega = \{(x_1, x_2) : \ 0 < x_1, x_2 < 1\}$ and assume $\Gamma_D = \{(x_1, x_2) : \ x_1 = 0\}$, $\Gamma_H = \partial\Omega \setminus \Gamma_D$. Again, we use $u_D(t) = h(t)$ for $x \in \Gamma_D$ with parameters $\alpha = 4$, $\beta = 2$, $\omega = 1$, and $\varepsilon = 0.1$. Figure 3 shows the profile of the solution u at time $t = 1.25$ with $\varepsilon = 0.1$.

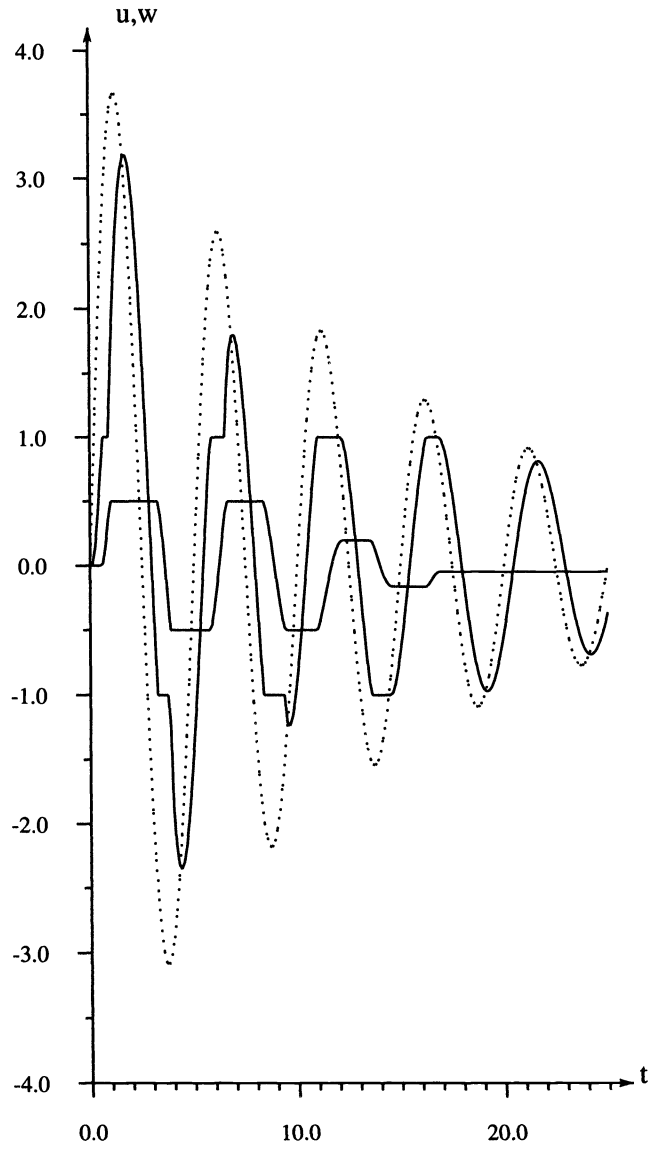


FIG. 1. *u and w as functions of t at x = 1 for Example 1 with ε = 0.*

5. A parabolic problem. We close with some remarks on a *parabolic* system obtained as an *additive* perturbation of $[\partial\zeta_a, \partial\zeta_b]$ instead of the composition $\mathbb{C}$ that was used in §3 to recover (1.1). The first is a corollary of Proposition 1.

COROLLARY 3. *Assume that ζ_a and ζ_b are given in Proposition 1 and (2.3.a)–(2.3.c) hold. For every $u_0 \in L^2(\Omega)$, $v_0 \in L^2(\Gamma)$ and $f \in L^2(0,T;L^2(\Omega))$, $g \in L^2(0,T;L^2(\Gamma))$, there is a unique solution $u \in C([0,T];L^2(\Omega))$, $v \in C([0,T];L^2(\Gamma))$*

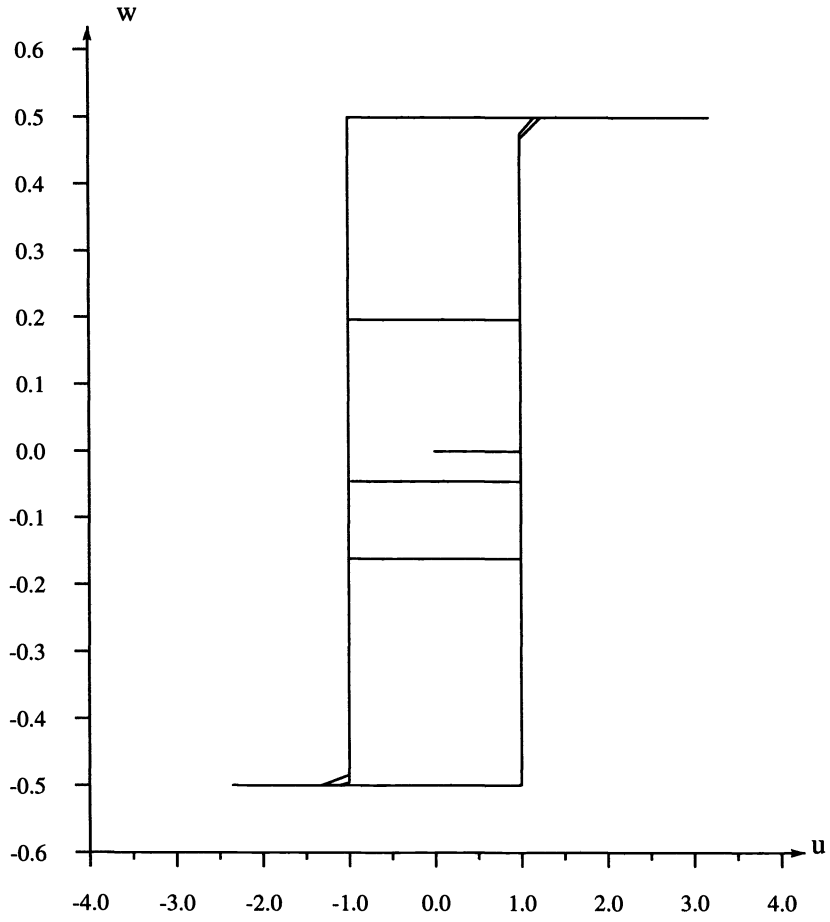


FIG. 2. *Relay: w versus u at $x = 1$ for Example 1 with $\varepsilon = 0$.*

of

$$\frac{\partial u}{\partial t} + a + Au = f \ , \quad a \in \partial\zeta_a(u) \ \ in \ \ L^2_{\mathrm{loc}}\big(0,T;L^2(\Omega)\big), \tag{5.1.a}$$

$$\frac{\partial v}{\partial t} + b + \partial_A u = g \ , \quad b \in \partial\zeta_b(v) \ , \ \ and \tag{5.1.b}$$

$$\partial_A u \in \partial\zeta_c(v - \gamma u) \ \ in \ \ L^2\big(0,T;L^2(\Gamma)\big) \ , \tag{5.1.c}$$

$$u(0) = u_0 \ \ in \ \ L^2(\Omega) \ , \quad v(0) = v_0 \ \ in \ \ L^2(\Gamma) \ . \tag{5.1.d}$$

Proof. Estimates (2.3.a)–(2.3.c) imply that ∂Z_a and ∂Z_b are defined everywhere, hence, by Corollary 2.7 of [7] we have $\partial Z = \partial Z_a + \partial Z_b + \partial Z_2$ in $L^2(\Omega) \times L^2(\Gamma)$. Then, (3.2) is the evolution generated by ∂Z. □

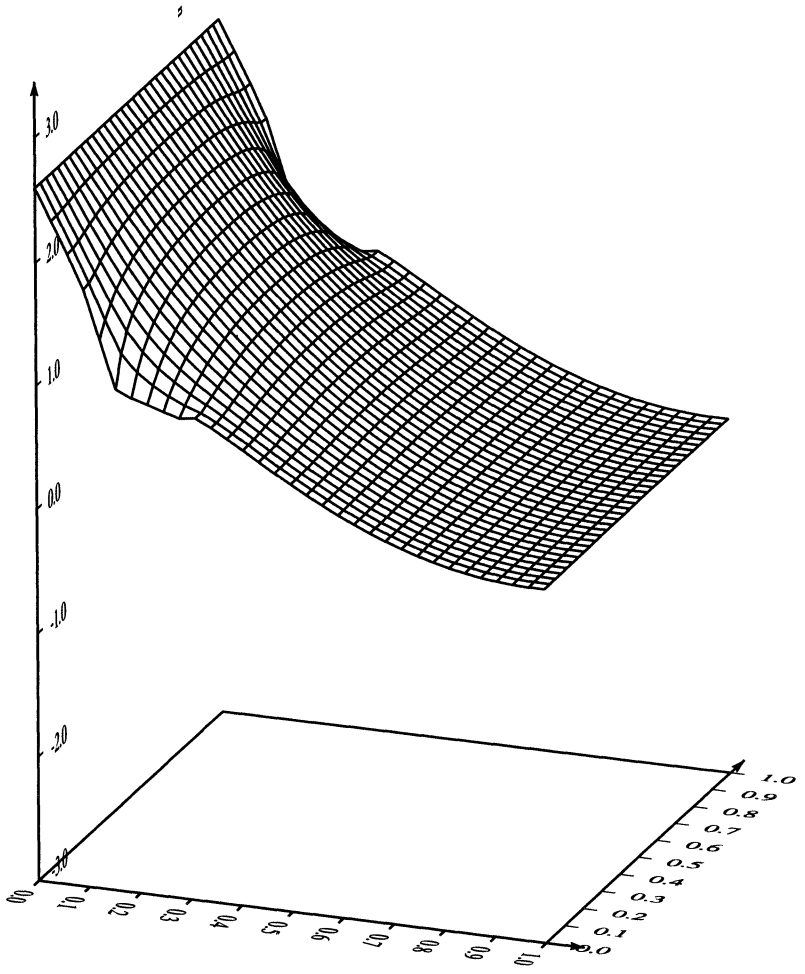


FIG. 3. *Profile at* $t = 1.25$ *for Example* 2 *with* $\varepsilon = 0.1$.

Such a subgradient induces a *parabolic regularizing effect* in the dynamics. Specifically, the solution of (3.2) is strongly differentiable and satisfies

$$u(t) \in D \quad \text{a.e.} \quad t \in (0,T) \ .$$

Also, we note from Theorem 1 that the stationary problem associated with (3.2) is well posed when two of the three parts of (2.11) hold.

The fact that $\mathbb{C}_2$ is ∂J-monotone for *any* J of the form (3.3) has many consequences for the special case of system (5.1) with $\zeta_a = \zeta_b = 0$. In particular, if

$(I+\varepsilon\mathbb{C}_2)([u_j, v_j]) = [f_j, g_j]$ for $j = 1, 2$ and $\varepsilon > 0$, then we have the *resolvent estimate*

$$J([u_1 - u_2,\, v_1 - v_2]) \le J([f_1 - f_2,\, g_1 - g_2]) \tag{5.2}$$

for any such J. Similar estimates hold for the evolution system, and any such J is a *Lyapunov function* for this special case of system (5.1). These lead to L^p-estimates and comparison theorems for solutions by taking appropriate choices of j. Finally, we note that $\partial(J + Z_2) = \partial J + \mathbb{C}_2$, and this leads to another parabolic case of (5.1).

COROLLARY 4. *Let j be given as in Proposition 2 of §3 and set $\zeta_a = \zeta_b = j$. Assume* (2.3.c) *holds. Then the result of Corollary 3 is valid.*

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] H. W. ALT, *On the thermostat problem*, Control Cybernet., 14 (1985), pp. 171–193; Math. Z., 183 (1983), pp. 311–341.

[3] PH. BENILAN, *Équations d'évolution dans un espace de Banach*, C.R. Acad. Sci. Paris, 274 (1972), pp. 47–50.

[4] PH. BENILAN, M.G. CRANDALL, AND P. E. SACKS, *Some L^1 existence and dependence results for semilinear elliptic equations under nonlinear boundary conditions*, Appl. Math. Optim., 17 (1988), pp. 203–224.

[5] H. BREZIS, *Monotonicity methods in Hilbert spaces and some application to nonlinear partial differential equations*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971.

[6] ———, *Problemes unilateraux*, J. Math. Pures Appl., 51 (1972), pp. 1–164.

[7] ———, *Opérateurs Maximaux Monotones et Semigroupes de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[8] H. BREZIS AND W. STRAUSS, *Semilinear elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 15–26.

[9] M.G. CRANDALL, *An introduction to evolution governed by accretive operators*, in Dynamical Systems—An International Symposium, L. Cesari, J. Hale, J. LaSalle, eds., Academic Press, New York, 1976, pp. 131–165.

[10] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North–Holland, Amsterdam, 1976.

[11] U. HORNUNG AND W. JÄGER, *Homogenization of reactive transport through porous media*, in International Conference on Differential Equations, Barcelona 1991, C. Perelló, C. Simó, and Solá-Morales, eds., World Scientific Publishing, Singapore, 1993, pp. 136–152.

[12] U. HORNUNG AND R. E. SHOWALTER, *PDE-Models with Hysteresis on the Boundary*, in Models of Hysteresis, Pitman Research Notes in Mathematics 286, A. Visintin, ed., Longman Scientific and Technical, Harlow, U.K., 1993, pp. 30–38.

[13] N. KENMOCHI AND A. VISINTIN, *Asymptotic stability for parabolic variational inequalities with hysteresis*, in Models of Hysteresis, Pitman Research Notes in Mathematics 286, A. Visintin, ed., Longman Scientific and Technical, Harlow, U.K., 1993, pp. 59–70.

[14] M. A. KRASNOSEL'SKII AND A. V. POKROVSKII, *Systems with Hysteresis*, Springer-Verlag, Berlin, 1989.

[15] T. D. LITTLE AND R. E. SHOWALTER, *The super-Stefan problem*, Internat. J. Engrg. Sci., to appear.

[16] ———, *Semilinear parabolic equations with Preisach hysteresis*, Differential Integral Equations, 7 (1994), pp. 1021–1040.

[17] J. W. MACKI, P. NISTRI, AND P. ZECCA, *Mathematical models for hysteresis*, SIAM Rev., 35 (1993), pp. 94–123.

[18] E. MAGENES, *On a Stefan problem on the boundary of a domain*, in Partial Differential Equations and Related Subjects, Pitman Research Notes in Mathematics 269, M. Miranda, ed., Longman Scientific and Technical, Harlow, U.K., 1992, pp. 209–226.

[19] I. D. MAYERGOYZ, *Mathematical Models of Hysteresis*, Springer-Verlag, Berlin, 1991.

[20] N. SAUER, *The Friedrichs extension of a pair of operators*, Quaestiones Math., 12 (1989),

pp. 239–249.

[21] R. E. SHOWALTER, *Hilbert Space Method for Partial Differential Equations*, Pitman, London, 1977.

[22] R. E. SHOWALTER AND N. J. WALKINGTON, *Diffusion of fluid in a fissured medium with microstructure*, SIAM J. Math. Anal., 22 (1991), pp. 1702–1722.

[23] A.J. VAN DER MERWE, *B-evolutions and Sobolev equations*, Appl. Anal., 29 (1988), pp. 91–105.

[24] A. VISINTIN, *Evolution problems with hysteresis in the source term*, SIAM J. Math. Anal., 17 (1986), pp. 1113–1138.

[25] ———, *Partial differential equations with hysteresis*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Pitman, Boston, 1987, pp. 226–232.

[26] ———, *Hysteresis and semigroups*, in Models of Hysteresis, Pitman Research Notes in Mathematics Series 286, A. Visintin, ed., Longman Scientific and Technical, Harlow, U.K., 1993, pp. 192–206.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 791–819, July 1995

002

REGULARITY FOR THE INTERFACES OF EVOLUTIONARY p-LAPLACIAN FUNCTIONS*

HI JUN CHOE† AND JONGSIK KIM‡

Abstract. The support of an evolutionary p-Laplacian function has a finite propagation speed. Here we consider various questions involving the interface, which is the boundary of the open set where the solution is positive. We especially study the initial behaviour and regularity of the interface. We find a necessary and sufficient condition for the interface to move. For the regularity questions we show that the interface is globally Hölder continuous employing the Harnack principle. Furthermore, we prove that the interface is Lipschitz continuous after a large time and globally Lipschitz continuous if the initial data satisfy certain nondegeneracy conditions.

Key words. interface, Hölder continuity of interface, Lipschitz continuity of interface

AMS subject classification. 35J

1. Introduction.

In this paper we consider the Cauchy problems for the evolutionary p-Laplace equation

$$u_t - \operatorname{div}\left(|\nabla u|^{p-2}\nabla u\right) = u_t - \Delta_p u = 0, \qquad p > 2 \tag{1}$$

in $R^n \times (0,\infty)$, $n \geq 1$, with a nonnegative continuous initial datum

$$u(x,0) = u_0(x)$$

of compact support. The main object is to study the interface Γ, which is the boundary of the open set where $u > 0$. These problems arise in geometry and non-Newtonian fluid mechanics (see [18] and [21]). Indeed, the analysis of the interface provides useful information for the propagation of the data.

Since $p > 2$, equation (1) is degenerate when $\nabla u = 0$. Hence the concept of classical solution is too restrictive. A weak solution of (1) is a function $u(x,t)$ such that for any $T > 0$,

$$\int_0^T \int_{R^n} u^2(x,t) + |\nabla u|^p \; dxdt < \infty$$

and

$$\int_0^T \int_{R^n} u\frac{\partial \phi}{\partial t} - |\nabla u|^{p-2}\nabla u \cdot \nabla \phi \; dxdt + \int_{R^n} u_0(x)\phi(x,0) \; dx \; = 0$$

for any continuously differentiable function ϕ with compact support in $R^n \times [0,T)$. The unique solvability of our Cauchy problem in $R^n \times (0,T)$ follows from Theorems 1 and 4 in [12].

* Received by the editors April 12, 1993; accepted for publication (in revised form) November 29, 1993.

† Department of Mathematics, Pohang Institute of Science and Technology, Pohang, Kyungbuk, Republic of Korea 790-600 (choe@posmath.postech.ac.kr). The research of this author was supported by Korea Science and Engineering Foundation, Global Analysis Research Center at Seoul National University, and by the Non Directed Research Fund, Korea Research Foundation, 1993.

‡ Department of Mathematics, Seoul National University, Seoul, Republic of Korea 151-742. The research of this author was supported by the Global Analysis Research Center at Seoul National University.

For the porous medium equation

$$v_t - \Delta(v^m) = 0, \quad m > 1,$$

various results for the interface are known. Assuming the initial datum v_0 has compact support, the interface consists of two parts—a moving boundary and a nonmoving boundary—and the support $\Omega(t) = \{x \in R^n : v(x,t) > 0\}$ is monotonically increasing. Furthermore, the regularity of interface has been investigated by many authors (see [1], [5]–[7], [17], etc.).

Here, following an argument similar to [5] and [7] we study the interface questions for the evolutionary p-Laplace equation (1). We define

$$\Lambda = \{(x,t) \in R^n \times [0,\infty) : u(x,t) > 0\},$$

$$\Omega(t) = \{x \in R^n : u(x,t) > 0\},$$

$$\Gamma(t) = \text{ the boundary of } \Omega(t),$$

and

$$\Gamma = \cup_{t \geq 0} \Gamma(t).$$

Hence $\Gamma(0)$ is the boundary of $\{x \in R^n : u_0 > 0\}$.

In §2 we consider the initial behaviour of the interface. When the initial datum u_0 satisfies

$$u_0(x) \geq c\,[\mathrm{dist}(x)]^{\frac{p}{p-2}}, \tag{2}$$

where $\mathrm{dist}(x) = \mathrm{distance}(x, \Gamma(0))$, then $\Omega(t)$ is initially strictly increasing. On the other hand we show that (2) is sufficient for strict monotonicity of the set $\Omega(t)$. In fact we show that if there is a supporting hyperplane P at $x_0 \in \Gamma(0)$ and

$$u_0(x) \leq c\,[\mathrm{dist}(x)]^{\gamma}, \quad \gamma > \frac{p}{p-2},$$

then there is a positive time $\tau > 0$ such that

$$u(x_0,t) = 0 \quad \text{for all } 0 < t < \tau,$$

and hence the interface does not move at x_0 for a short period. There are corresponding results for porous medium equations (see [17] for one dimension and [5] for a higher dimension). Finally we find an integral expression which describes the pointwise behaviour of the interface. In fact we use the Harnack-type inequality and an integral estimate by DiBenedetto and Herrero [12] and obtain a necessary and sufficient condition for moving point.

In §3, from the Harnack principle and the maximum principle we prove Hölder regularity of the interface. First we show that the interface consists of two parts—a moving part Γ_1 and a nonmoving part Γ_2. In particular, if the interface Γ contains a vertical line segment $\sigma = \{(x,t) : x = x_0, t_0 < t < t_1\}, 0 < t_0 < t_1$, then the entire segment $\{(x,t) : x = x_0, 0 < t < t_1\}$ belongs to Γ. Following an iteration method, we show that if $(x_0,t_0) \in \Gamma$ does not lie on a vertical line segment belonging to Γ, then

$\Gamma \cap \{t = \tau\}$ increases at a rate $\geq (\tau - t_0)^\mu$ for $\tau > t_0, \tau - t_0$ small, $|x - x_0|$ small, for some $\mu > 1$. Solutions of porous medium equations show the same behaviour (see Theorem 3.2 in [5]). Therefore the Hölder continuity of the interface follows. Here, the Harnack principle is a main tool in showing the above results.

In §4, following the idea of [7], we prove Lipschitz regularity of the interface after a large time. In fact, Theorem 6 is observed in [15]. Here, the asymptotic behaviour of the solution is crucial in estimating the Lipschitz norm of the solution after a large time. On the other hand, from the reflection method we find a monotonicity property of the solution. Indeed, the same property (see Proposition 1.5 in [7]) is used for the porous medium equation. Hence the interface is representable as a function in polar coordinates. Finally, by considering a directional derivative we show that the interface is Lipschitz after a large time.

In §5, employing the method of [7], we show the interface is Lipschitz continuous if the initial datum satisfies a nondegeneracy condition. A similar result for the porous medium equation is proved by Caffarelli, Vazquez, and Wolanski [7]. We show that a directional derivative of $v = u^{(p-2)/(p-1)}$ is positive and hence the interface is Lipschitz continuous. Here, the observation by Esteban and Vazquez [15] that v satisfies

$$v_t = \left(\frac{p-1}{p-2}\right)^{p-2} v\Delta_p v + \left(\frac{p-1}{p-2}\right)^{p-1} |\nabla v|^p$$

is rather crucial.

2. Initial behaviour of the interface. In this section we study the initial growth rate of the interface. The local behaviour of the initial datum near $\Gamma(0)$ is crucial in showing that $\Gamma(t)$ is increasing.

Suppose x_0 is on $\Gamma(0)$. Following an argument similar to Knerr [17], it is shown that the Hölder exponent of ∇u is critical in the study of the behaviour of the interface. Indeed, constructing a sequence of subsolutions near interface, we prove that if

$$u_0(x) \geq [\mathrm{dist}(x)]^{\frac{\gamma}{p-2}}$$

for some $\gamma < p$, then the interface is moving near x_0. Otherwise, the interface does not move for a short period.

LEMMA 1. *Let $\Gamma(0)$ be of C^2. Suppose that $B_R(0) \subset \Omega(0)$ and $\overline{B_R(0)} \cap \Gamma(0) = \{x_0\}$. Furthermore, we assume that*

$$u_0(x) \geq [\mathrm{dist}(x)]^{\frac{\gamma}{p-2}} \tag{3}$$

for some $\gamma < p$. Then we have

$$u(x_0, t) > 0$$

for all $t > 0$.

Proof. Without loss of generality we may assume that

$$p - 1 < \gamma < p.$$

Let $g(r) = (R - r)^\eta$, where $\eta = \frac{\gamma}{p-1}$. Note that $\eta > 1$; this is crucial for the following argument. Since we are assuming

$$\overline{B_R(0)} \cap \Gamma(0) = \{x_0\},$$

we see that

$$|x_0| = R.$$

Now we choose a sufficiently close to R such that

$$\frac{n-1+\frac{1}{2}(\eta-1)(p-1)}{n-1+(\eta-1)(p-1)}R < a < R.$$

We find s satisfying

$$g'(s) = -\eta(R-s)^{\eta-1} = \frac{1}{2}\frac{g(a)}{a-R}$$

and set $\tilde{x}$ as

$$\tilde{x} = s\frac{x_0}{|x_0|}.$$

Since g is a convex function and $g'(R) = 0$, it is always possible to choose such s. We set $s_1 = s$. Define recursively

$$s_{k+1} = \frac{s_k + R}{2}$$

for $k = 1, 2, 3, \ldots$ and

$$\tilde{x}_k = s_k\frac{x_0}{|x_0|}.$$

We define a supporting cone T_k at $(\tilde{x}_k, g(|\tilde{x}_k|))$ as

$$T_k(x) = g(s_k) + \eta(R-s_k)^{\eta-1}(s_k - |x|).$$

We see that

$$T_k(0) = g(s_k) + \eta(R-s_k)^{\eta-1}s_k.$$

Moreover, when

$$|x| = s_k + \frac{g(s_k)}{\eta(R-s_k)^{\eta-1}} = s_k + \frac{1}{\eta}(R-s_k) \equiv \tilde{s}_k,$$

we find that

$$T_k(x) = 0.$$

Since $g(s)$ is convex near R, we conclude that

$$T_k(x) \le g(|x|)$$

for all $x \in B_R(0)$. We also observe that $T_k(0)$ decreases monotonically as k goes to ∞.

Now we are ready to construct comparison functions that are subsolutions to (1). Since the maximum principle holds for solutions of (1), we can assume

$$u_0(x) = u(x,0) = \left[(R-|x|)^+\right]^\theta, \quad \theta = \frac{\gamma}{p-2}$$

without loss of generality. From a result of Lieberman (see [20]), if ∇u_0 is Hölder continuous, then ∇u is Hölder continuous up to $t = 0$. Hence it follows that

$$\begin{aligned} u_t(x,0) &= \operatorname{div}\left(|\nabla u|^{p-2}\nabla u\right) \\ &= \theta^{p-1}\left(R-|x|\right)^{(\theta-1)(p-1)-1}\left[(\theta-1)(p-1)-\frac{n-1}{|x|}\left(R-|x|\right)\right] \geq 0 \end{aligned}$$

if

$$\frac{n-1}{(\theta-1)(p-1)+n-1}R \leq |x| \leq R.$$

Recall from the choice of a that

$$\frac{n-1}{(\theta-1)(p-1)+n-1}R \leq a \leq R.$$

So we see that there exists $\tau > 0$ such that

$$u_t(\bar{x},t) > 0 \quad \text{for all} \quad t \in [0,\tau),$$

where $\bar{x} = a\frac{x_0}{|x_0|}$.

Now we consider a family of functions $\{f_k(x,t)\}$ given by

$$f_k(x,t) = \alpha_k^p t + \alpha_k(\tilde{s}_k - |x|) \quad \text{if } |x| < \tilde{s}_k + \alpha_k^{p-1}t$$

and

$$f_k(x,t) = 0 \quad \text{otherwise,}$$

where $\alpha_k = \eta(R-s_k)^{\eta-1}$. Note that $f_k(x,0) = T_k(x)$. We define

$$u_k(x,t) = \left[f_k(x,t)\right]^q,$$

where $q = \frac{p-1}{p-2}$. Then, after suitable calculation, we have that

$$\left(u_k(x,t)\right)_t = q\alpha_k^p\left[\alpha_k^p t + \alpha_k(\tilde{s}_k - |x|)\right]^{q-1}$$

and

$$\begin{aligned} \operatorname{div}\left(|\nabla u_k|^{p-2}\nabla u_k\right) = {} & q^p\alpha_k^p\left[\alpha_k^p t + \alpha_k(\tilde{s}_k - |x|)\right]^{q-1} \\ & -q^{p-1}\alpha_k^{p-1}\left[\alpha_k^p t + \alpha_k(\tilde{s}_k - |x|)\right]^q \frac{n-1}{|x|}. \end{aligned}$$

Thus we have

$$\begin{aligned} &\left(u_k(x,t)\right)_t - \operatorname{div}\left(|\nabla u_k|^{p-2}\nabla u_k\right) \\ &= \alpha_k^p\left[\alpha_k^p t + \alpha_k(\tilde{s}_k - |x|)\right]^{q-1}\left(q - q^p + q^{p-1}\frac{n-1}{\alpha_k|x|}\left[\alpha_k^p t + \alpha_k(\tilde{s}_k - |x|)\right]\right). \end{aligned}$$

Therefore, if

$$q - q^p + q^{p-1}\frac{n-1}{|x|}\left[\alpha_k^{p-1}t + (\tilde{s}_k - |x|)\right] \leq 0,$$

that is,

$$|x| \geq \frac{q^{p-1}(n-1)(\alpha_k^{p-1}t + \tilde{s}_k)}{q^p + q^{p-1}(n-1) - q},$$

then u_k is a subsolution to (1) and

$$(u_k(x,t))_t - \operatorname{div}\left(|\nabla u_k|^{p-2}\nabla u_k\right) \leq 0.$$

We note that

$$u_k(x_0,t) = 0 \quad \text{for all } 0 \leq t \leq \tau_k$$

with

$$\tau_k = \frac{1}{\alpha_k^{p-1}}(R - \tilde{s}_k).$$

Moreover, we find that

$$\begin{aligned}\lim_{k\to\infty} \tau_k &= \lim_{k\to\infty} \frac{1}{\alpha_k^{p-1}}(R-\tilde{s}_k) \leq \lim_{k\to\infty} \frac{R-\tilde{s}_k}{[\eta(R-\tilde{s}_k)^{\eta-1}]^{p-1}} \\ &= \lim_{k\to\infty} \frac{1}{\eta^{p-1}}(R-\tilde{s}_k)^{p-\eta(p-1)} = 0,\end{aligned}$$

since

$$R > \tilde{s}_k \geq s_k \quad \text{and} \quad p - \eta(p-1) > 0.$$

Now, from a direct computation it is rather simple to see that

$$f_k(\bar{x}, \tau_k) = \alpha_k(R - |\tilde{x}|) = \alpha_k(R-a),$$

and since $\alpha_k < \alpha_1$ for all $k \geq 2$, it follows from the choice of s_1 that

$$f_k(\bar{x}, \tau_k) < \alpha_1(R-a) = \frac{1}{2}g(a) = \frac{1}{2}g(|\bar{x}|).$$

Now, recall that

$$q = \frac{p-1}{p-2} > 1$$

and

$$u_t(\bar{x}, t) > 0 \quad \text{for all } 0 < t < \tau.$$

Therefore we conclude that for $R \leq 1$,

$$u_k^{\frac{1}{q}}(\bar{x}, t) = f_k(\bar{x}, t) \leq g(|\bar{x}|) \leq u^{\frac{1}{q}}(\bar{x}, t)$$

for all $0 \leq t < \tau$, since

$$u_t(\bar{x}, t) > 0 \quad \text{for all } 0 < t < \tau.$$

Recall that u_k is a subsolution to (1). Consequently, from the comparison principle,

$$u_k^{\frac{1}{q}}(x_0,t) = f_k(x_0,t) \le u^{\frac{1}{q}}(x_0,t)$$

for all $\tau_k < t < \tau$, and since

$$\lim_{k\to 0} \tau_k = 0,$$

we conclude that

$$u(x_0,t) > 0$$

for all $0 < t < \tau$. □

Now we prove a lemma that is a converse of Lemma 1. The growth condition (2) is almost sufficient for showing that the interface is moving at $(x_0, 0)$.

LEMMA 2. *Suppose that $x_0 \in \Gamma(0)$ and there is a supporting hyperplane P in R^n such that $x_0 \in P$ and $\Omega(0)$ lies completely in one side of P. If*

$$u_0(x) \le [\operatorname{dist}(x,\Gamma(0))]^{\frac{p}{p-2}}, \tag{4}$$

then there is a small positive constant $\tau > 0$ such that

$$u(x_0,t) = 0$$

for all $0 \le t \le \tau$.

Proof. We find a supersolution bounding u. Since the partial differential equation (1) is invariant under translation and rotation for the space variable x, we assume that x_0 is the origin, $P = \{x \in R^n : x_n = 0\}$, and $\Omega(0)$ is contained in the upper half space. We notate $x = (x', x_n)$ and hence $x' = (x_1, x_2, \ldots, x_{n-1})$. Since $\Omega(0)$ lies in the upper half space, we have for each $x \in \Omega(0)$,

$$\operatorname{dist}(x,\Gamma(0)) \le x_n$$

and from the assumption (4),

$$u_0(x) \le x_n^{\frac{p}{p-2}}.$$

We choose M so that

$$M \ge u(x,t)$$

for $0 \le t \le T_0$, where T_0 is a fixed positive time. We define $d = \operatorname{diam}(\Omega(0))$; then $u_0(x', x_n) = 0$ on $\partial B_d'(x') \times (0,d)$ and $u_0(x', 0) = 0$, where $B_d'(x')$ is the $(n-1)$-dimensional ball centered at x' with radius d. By direct computation we have that

$$w = b\left(\frac{x_n^p}{T-t}\right)^{\frac{1}{p-2}}$$

is a solution to (1), where

$$b = \left(\frac{p-2}{p}\right)^q \frac{1}{[2(p-1)]^{p-2}}.$$

Fix $\delta > 0$. Now we take T so small that

$$b\left(\frac{\delta^p}{T}\right)^{\frac{1}{p-2}} \geq 2M.$$

Hence we obtain

$$w(x,0) \geq u_0(x)$$

and

$$w(x,t) \geq M$$

for all $(x,t) \in B_d'(x') \times \{\delta\} \times (0,T)$. Therefore, from the comparison principle we conclude that

$$w(x,t) \geq u(x,t)$$

for all $(x,t) \in B_d'(x') \times (0,\delta) \times (0,T)$ and, in particular,

$$w(0,t) = u(0,t) = 0$$

for all $0 \leq t < T$. □

Now we find an integral expression which describes the initial behaviour of the interface. From the Harnack-type inequalities we prove the following theorem which implies Lemmas 1 and 2.

THEOREM 1. *Define*

$$I(x) = \sup_R R^{-n-\frac{p}{p-2}} \int_{B_R(x)} u_0(y)dy.$$

Given that $x \in R^n$, we have $u(x,t) > 0$ for all $t > 0$ if and only if $I(x) = \infty$, that is,

$$\cap_{t>0}\Omega(t) = \{x : I(x) = \infty\}.$$

Moreover, there exists a constant $c = c(n,p) > 0$ such that $u(x,t) = 0$ for every (x,t) such that

$$0 < t < cI^{2-p}(x).$$

Proof. Suppose $I(x) = \infty$. From the Harnack principle (see Corollary 1 in [9]) we have that

$$R^{-n-\frac{p}{p-2}} \int_{B_R(x)} u_0(x)dx \leq c\left(t^{-\frac{1}{p-2}} + t^{\frac{n}{p}} R^{-n-\frac{p}{p-2}} u(x,t)^{\frac{n(p-2)+p}{p}}\right). \tag{5}$$

Hence, if $u(x,t) = 0$ for some $t > 0$, then

$$I(x) \leq ct^{-\frac{1}{p-2}};$$

this contradicts $I(x) = \infty$.

Now we assume $I(x) < \infty$. From Theorem 1 in [12] we know that

$$\sup_{B_\rho} u(x,t) \leq ct^{-\frac{n}{\kappa}} \rho^{\frac{p}{p-2}} (I_R(x))^{\frac{p}{\kappa}}$$

for all $\rho > R > 0$ and $0 < t < T(R)$, where

$$I_R(x) = \sup_{\rho > R} \rho^{-n-\frac{p}{p-2}} \int_{B_\rho(x)} u_0(y)dy$$

and

$$T(R) = c\,[I(x)]^{-(p-2)}.$$

Taking $R \to 0$ we are set. □

3. Hölder continuity of the interface. In this section we show that the interface is a Hölder continuous graph as a function of x. A main tool is the Harnack principle, which is proved by DiBenedetto [9].

THEOREM 2 (Harnack principle by DiBenedetto). *Let u be a nonnegative weak solution of* (1). *Let $(x_0, t_0) \in R^n \times (0, T)$ and $B_R(x_0)$ be the ball of radius R centered at x_0. We assume $u(x_0, t_0) > 0$. Then there are constants c_0 and c_1 depending only on n and p such that*

$$u(x_0, t_0) \le c_0 \inf_{x \in B_R(x_0)} u(x, t_0 + \theta), \tag{6}$$

where

$$\theta = \frac{c_1 R^p}{[u(x_0, t_0)]^{p-2}},$$

provided $t_0 \ge \theta$.

For the proof of the Harnack principle, the fundamental solution $\Phi_{k,\rho}$ to (1) plays a central role (see [9]):

$$\Phi_{k,\rho}(x, t; \bar{x}, \bar{t}) = k\rho^n S(t)^{-\frac{n}{\kappa}} \left[1 - \left(\frac{|x - \bar{x}|}{S(t)^{\frac{1}{\kappa}}}\right)^{\frac{p}{p-1}}\right]_+^{\frac{p-1}{p-2}},$$

where

$$S(t) = (\gamma_0(n,p)k^{p-2}\rho^{n(p-2)}(t - \bar{t}) + \rho^\kappa), \quad t \ge \bar{t},$$

$$\gamma_0(n,p) = \kappa\left(\frac{p}{p-2}\right)^{p-1}, \quad \kappa = n(p-2) + p.$$

Considering the above fundamental solutions, we can show that if $\overline{\Omega(0)}$ is compact, then $\overline{\Omega(t)}$ is compact for all $t \ge 0$. Moreover, an integral estimate independent of scaling follows from the Harnack principle and we omit the proof (compare with Corollary 1 in [9]).

LEMMA 3. *For all $R, \theta > 0$ such that $Q_{2R}(\theta) \subset R^n \times (0, \infty)$, the following holds:*

$$\fint_{B_R(x_0)} u^p(x, t_0)dx \le B\left(\left(\frac{R^p}{\theta}\right)^{\frac{p}{p-2}} + \left(\frac{\theta}{R^p}\right)^n \left[\inf_{x \in B_R(x_0)} u(x_0, t_0 + \theta)\right]^\kappa\right) \tag{7}$$

for some positive constant B depending only on n and p, where the cylinder $Q_{2R}(\theta)$ is defined by

$$Q_{2R}(\theta) = B_{2R}(x_0) \times (t_0 - \theta, t_0 + \theta),$$

and

$$\kappa = n(p-2) + p.$$

For the Harnack principle, the condition that $p > 2$ is rather critical as the following example shows. Let $\frac{2n}{n+2} < m < 2$ and u be the solution to a Dirichlet problem

$$u_t - \operatorname{div}\left(|\nabla u|^{m-2}\nabla u\right) = 0 \quad \text{in } B_R(0) \times (0,\infty)$$

with initial boundary condition $u(x,0) = u_0(x) \geq 0$ for all $x \in B_R(0)$ and lateral boundary condition $u(x,t) = 0$ for all $(x,t) \in \partial B_R(0) \times (0,\infty)$. Then there is a finite time T depending on $||u_0||_{L^2}$ such that

$$u(x,t) = 0 \quad \text{for all } (x,t) \in B_R(0) \times (T,\infty).$$

So we cannot expect the Harnack principle of the form (6).

The monotonicity of the interface follows immediately from the Harnack principle.

THEOREM 3. *$\Omega(t)$ is monotonically increasing, that is,*

$$\Omega(t_1) \subset \Omega(t_2) \quad \text{if } 0 < t_1 \leq t_2.$$

Proof. Let $x_0 \in \Omega(t_1)$; then

$$u(x_0,t_1) > 0$$

and there exists a small ball $B_{R_0}(x_0) \subset \Omega(t_1)$. Define

$$\theta_0 = \frac{c_1 R_0^p}{[u(x_0,t_1)]^{p-2}}$$

and assume R_0 is sufficiently small so that

$$t_1 \geq \theta_0.$$

Hence, from the Harnack principle we have

$$u(x_0,t_1) \leq c_0 \inf_{x \in B_R(x_0)} u(x, t_1 + \theta)$$

for all $R < R_0$, where

$$\theta = \frac{c_1 R^p}{[u(x_0,t_1)]^{p-2}}.$$

Now, if

$$t_2 < t_1 + \theta_0,$$

then taking R sufficiently small we have

$$u(x_0,t_1) \leq c_0 u(x, t_1 + h) \tag{8}$$

for all $0 < h < \theta_0$. We observe that since $p > 2$, θ_0 goes to ∞ as $u(x_0,t_1)$ goes to 0. By the maximum principle, u is bounded and $\theta_0 > \varepsilon$ for some fixed positive number ε. Therefore, if

$$t_2 \geq t_1 + \theta_0,$$

we can iterate (8) and obtain

$$u(x_0, t_1) \leq c_0^k u(x, t_2)$$

for some k. Therefore

$$u(x_0, t_2) > 0$$

and $x_0 \in \Omega(t_2)$. □

Indeed following an argument of Benilan and Crandall [3] and [11], Theorem 3 can be proved without referring to the Harnack principle. We can show that the unique solution v with initial datum $v(x,0) = \lambda^{\frac{1}{p-2}} u_0(x), \lambda > 0$ is given by

$$v(x,t) = \lambda^{\frac{1}{p-2}} u(x, \lambda t).$$

If $\lambda > 1$, then $v(x,0) \geq u(x,0)$. Hence, from the comparison principle we have $u(x,t) \leq v(x,t)$ for all $(x,t) \in \boldsymbol{R}^n \times (0,\infty)$. Choosing $\lambda = 1 + \frac{h}{t}$ for a small positive number h, we obtain

$$u(x, t+h) - u(x,t) = u(x, \lambda t) - u(x,t) = \lambda^{\frac{1}{2-p}} v(x,t) - u(x,t)$$

$$\geq \left(\lambda^{\frac{1}{2-p}} - 1\right) u(x,t).$$

Dividing by h and sending h to 0, we conclude that

$$u_t \geq -\frac{1}{p-2} \frac{u}{t};$$

this implies Theorem 3.

Now, considering Theorem 1 we cannot expect that $\Omega(t)$ is strictly increasing.

Define a cylinder $Q_R^h(x,t)$ by

$$Q_R^h(x,t) = B_R(x) \times (t, t+h).$$

Following a Moser-type iteration method we have a local maximum principle.

LEMMA 4. *Suppose $u(x,t_0) = 0$ for all $x \in B_{R_0}(x_0)$. Then we have*

$$\sup_{Q^h_{\frac{R_0}{2}}(x_0,t_0)} u \leq c \left(\frac{h}{R_0^p}\right)^{\frac{1}{2}} \left(\fint_{Q^h_{R_0}(x_0,t_0)} u^p \, dz \right)^{\frac{1}{2}} \tag{9}$$

for some c depending only on n and p, where we defined

$$\fint_A u dz = \frac{1}{|A|} \int_A u dz.$$

Proof. We omit the expression (x_0, t_0) of the generic point in various terms if it is clear. Let $0 < \rho < R < R_0$ and let η be a cutoff function such that $\eta \equiv 1$ in B_ρ and let $\eta \in C_0^\infty(B_R)$. Considering a suitable approximation, we can take $u^{\alpha+1}\eta^p$ as a test function for all $\alpha \geq 0$. Hence we have

$$\frac{1}{\alpha+2} \int \frac{d}{dt} \left(u^{\alpha+2}\right) \eta^p \, dxdt + \int |\nabla u|^{p-2} \nabla u \cdot \nabla \left(u^{\alpha+1}\eta^p\right) \, dxdt = 0,$$

and from the assumption that $u \equiv 0$ on $B_R \times \{t_0\}$ we obtain

$$\frac{1}{\alpha+2}\sup_t \int_{B_R} u^{\alpha+2}\eta^p \, dx + (\alpha+1)\int_{Q_R^h} |\nabla u|^p u^\alpha \eta^p \, dxdt$$

$$\leq \frac{c}{\alpha+1}\int_{Q_R^h} u^{\alpha+p}|\nabla \eta|^p \, dxdt$$

for some c. We assume $n \geq 2$. The case $n = 2$ can be proved similarly. From Sobolev's imbedding theorem and the Hölder inequality we obtain

$$\begin{aligned}\int_{Q_\rho^h} u^{(\alpha+p)+\frac{2}{n}(\alpha+2)} \, dxdt &= \int_{t_0}^{t_0+h}\int_{B_\rho} u^{(\alpha+p)+\frac{2}{n}(\alpha+2)} \, dxdt \\ &\leq \left[\sup_t \int_{B_\rho} u^{\alpha+2} \, dx\right]^{\frac{2}{n}} \int_{t_0}^{t_0+h}\left[\int_{B_\rho} u^{(\alpha+p)\frac{n}{n-2}} \, dx\right]^{\frac{n-2}{n}} dt \\ &\leq \left[\sup_t \int_{B_\rho} u^{\alpha+2} \, dx\right]^{\frac{2}{n}} \int_{t_0}^{t_0+h}\int_{B_R} \left|\nabla\left(u^{\frac{\alpha+p}{2}}\eta\right)\right|^2 \, dxdt\end{aligned}$$

for some c depending only on α, n, and p. We also have

$$\begin{aligned}&\int_{t_0}^{t_0+h}\int_{B_R} \left|\nabla\left(u^{\frac{\alpha+p}{2}}\eta\right)\right|^2 \, dxdt \\ &\quad \leq c\int_{Q_R^h} |u|^{\alpha+p-2}|\nabla u|^2\eta^2 \, dz + \frac{c}{(R-\rho)^2}\int_{Q_R^h} |u|^{\alpha+p} \, dz\end{aligned}$$

and from the Hölder inequality

$$\int_{Q_R^h} |u|^{\alpha+p-2}|\nabla u|^2\eta^2 \, dz \leq c\left[\int_{Q_R^h} |\nabla u|^p u^\alpha \eta^p \, dz\right]^{\frac{2}{p}}\left[\int_{Q_R^h} |u|^{\alpha+p} \, dz\right]^{1-\frac{2}{p}}$$

for some c depending only on α, n, and p. Combining all these together we have

$$\int_{Q_\rho^h} u^{(\alpha+p)+\frac{2}{n}(\alpha+2)} \, dz \leq \frac{c}{(R-\rho)^{2+\frac{2p}{n}}}\left[\int_{Q_R^h} |u|^{\alpha+p} \, dz\right]^{1+\frac{2}{n}} \tag{10}$$

for some c depending only on α, n, and p.

We iterate (10). Define

$$\alpha_{\nu+1} = \left(1+\frac{2}{n}\right)\alpha_\nu + \frac{4}{n}$$

with $\alpha_0 = 0$. Then we note that

$$\alpha_\nu = 2(\beta^\nu - 1),$$

where $\beta = 1 + \frac{2}{n}$. Let $\rho_\nu = (R_0/2)(1+2^{-\nu})$, $\nu = 0, 1, 2, \ldots$, and

$$Q_\nu = Q_{\rho_\nu}^h.$$

We set

$$\phi_\nu = \fint_{Q_\nu} u^{\alpha_\nu + p}\, dz.$$

Then we can write (10) as

$$\begin{aligned}(11)\qquad \phi_{\nu+1} &= \fint_{Q_{\nu+1}} u^{\alpha_{\nu+1}+p}\, dz \\ &\le \left(\frac{h}{R_0^p}\right)^{\frac{2}{n}} c^{\nu+1} \left(\fint_{Q_\nu} u^{\alpha_\nu+p}\, dz\right)^{\beta} = c^{\nu+1}\left(\frac{h}{R_0^p}\right)^{\frac{2}{n}} \phi_\nu^\beta.\end{aligned}$$

Set

$$\gamma = \left(\frac{h}{R_0^p}\right)^{\frac{2}{n}}.$$

Iterating (11) we obtain

$$\phi_\nu \le c^\nu \gamma \phi_{\nu-1}^\beta \le c^\nu \gamma \left[c^{\nu-1}\gamma \phi_{\nu-2}^\beta\right]^\beta$$

$$= c^{\nu+(\nu-1)\beta} \gamma^{1+\beta} \phi_{\nu-2}^{\beta^2}$$

$$\vdots$$

$$\le c^{\nu+(\nu-1)\beta+\cdots+\beta^{\nu-1}} \gamma^{1+\beta+\beta^2+\cdots+\beta^{\nu-1}} \phi_0^{\beta^\nu}$$

and

$$[\phi_\nu]^{\frac{1}{\beta^\nu}} \le c\gamma^{\frac{\beta^\nu - 1}{\beta-1}\frac{1}{\beta^\nu}} \phi_0.$$

Sending ν to ∞, we find that

$$\lim_{\nu\to\infty} \phi_\nu^{\frac{1}{\beta^\nu}} = \sup_{Q^h_{\frac{R}{2}}} u^2$$

and

$$\lim_{\nu\to\infty} \gamma^{\frac{\beta^\nu-1}{\beta-1}\frac{1}{\beta^\nu}} = \gamma^{\frac{1}{\beta-1}} = \frac{h}{R_0^p}.$$

Therefore we conclude that

$$\sup_{Q^h_{\frac{R_0}{2}}(x_0,t_0)} u \le c\left(\frac{h}{R_0^p}\right)^{\frac{1}{2}} \left(\fint_{Q^h_{R_0}(x_0,t_0)} u^p\, dz\right)^{\frac{1}{2}}$$

for some c depending only on n and p. □

Under the assumption of Lemma 4 that $u(x,t_0) \equiv 0$ in $B_{R_0}(x_0)$, it is shown that if the input of the total mass is small, the speed of propagation of the mass is small. Here the Harnack principle is a main tool.

LEMMA 5. *Suppose that $u(x,t_0) \equiv 0$ in $B_R(x_0)$. Let $h < \frac{t_0}{2}$. There exists a large constant c such that if*

$$\fint_{B_R(x_0)} u^p(x,t_0+h)\,dx \le \frac{1}{c}\left(\frac{R^p}{h}\right)^{\frac{p}{p-2}},$$

then

$$u(x,t) \equiv 0$$

in $B_{\frac{R}{4}}(x_0) \times (t_0, t_0+h)$.

Proof. With the maximum principle (see Lemma 4) we have

$$\sup_{Q^h_{\frac{R}{2}}(x_0,t_0)} u \le c\left(\frac{h}{R^p}\right)^{\frac{1}{2}} \sup_{Q^h_R(x_0,t_0)} u^{\frac{p}{2}}.$$

Let $x \in B_{\frac{R}{4}}(x_0)$; then $B_{\frac{R}{4}}(x) \subset B_{\frac{R}{2}}(x_0)$. Set $R_k = \frac{R}{2^k}, k = 1,2,3,\ldots$; then we have

$$M_{k+1} \le c\left(2^{kp}\frac{h}{R^p}\right)^{\frac{1}{2}} M_k^{\frac{p}{2}},$$

where $M_k = \sup_{Q^h_{R/2^k}(x,t_0)} u$. Since we are assuming $\frac{p}{2} > 1$, we obtain

$$M_k \to 0 \quad \text{if } M_1 < \frac{1}{c\left(\frac{h}{R^p}\right)^{\frac{1}{p-2}}}$$

for some c. In other words if

$$\sup_{Q^h_{\frac{R}{2}}(x_0,t_0)} u \le \frac{1}{c}\left(\frac{R^p}{h}\right)^{\frac{1}{p-2}} \tag{12}$$

for some large c, then

$$u(x,t) = 0$$

for all $t_0 < t < t_0 + h$ and all $x \in B_{R/4}(x_0)$.

Now we show that (12) is true if

$$\fint_{B_R(x_0)} u^p(x,t_0+h)\,dx \le \frac{1}{c}\left(\frac{R^p}{h}\right)^{\frac{p}{p-2}}$$

for some large c. From the Harnack principle we get

$$u(x,t) \le cu(x,t_0+h) \tag{13}$$

for all $t_0 < t < t_0 + h$ and all $x \in B_R$. Considering the maximum principle (9) and the Harnack principle (13) we have

$$\begin{aligned}\sup_{Q^h_{\frac{R}{2}}(x_0,t_0)} u &\le c\left(\frac{h}{R^p}\right)^{\frac{1}{2}} \left(\fint_{Q^h_R(x_0,t_0)} u^p \, dz\right)^{\frac{1}{2}} \\ &\le c\left(\frac{h}{R^p}\right)^{\frac{1}{2}} \left[\frac{1}{h}\int_{t_0}^{t_0+h} dt \fint_{B_R(x_0)} u^p(x,t)\, dx\right]^{\frac{1}{2}} \\ &\le c\left(\frac{h}{R^p}\right)^{\frac{1}{2}} \left[\fint_{B_R(x_0)} u^p(x,t_0+h)\, dx\right]^{\frac{1}{2}}\end{aligned}$$

for some c. Hence, if

$$\fint_{B_R(x_0)} u^p(x,t_0+h)\, dx \le \frac{1}{c}\left(\frac{R^p}{h}\right)^{\frac{p}{p-2}}$$

for some large c, we obtain

$$M_1 \le \sup_{Q^h_{\frac{R}{2}}(x_0,t_0)} u \le \frac{1}{c}\left(\frac{h}{R^p}\right)^{\frac{1}{2}} \left(\frac{R^p}{h}\right)^{\frac{p}{2(p-2)}} = \frac{1}{c}\left(\frac{R^p}{h}\right)^{\frac{1}{p-2}}.$$

This implies that

$$u(x,t) = 0$$

for all $(x,t) \in B_{R/4} \times (t_0, t_0+h)$ and completes the proof. □

As in the case of porous medium equations we prove that the interface consists of a moving part and a nonmoving part. We prove this by contradiction. We refer to [5] for the porous medium equations.

LEMMA 6. *Define* $\Gamma_1 = \{(x,t) \in \Gamma : \{(x,s) : s \ge 0\} \cap \Gamma = \{(x,t)\}\}$ *and* $\Gamma_2 = \{(x,t) \in \Gamma : t > 0$ *and* $\{(x,s) : s > 0\} \cap \Gamma = \{(x,s) : 0 \le s \le t\}\}$. *Then* $\Gamma_1 \cup \Gamma_2 = \Gamma$.

Proof. Suppose the assertion is not true. Then, for some $(x_0,t_0) \in \Gamma$ there exist t_1 and t_2 with $0 < t_1 < t_2 < t_0$ such that

$$u(x,t_1) = 0 \quad \text{for } x \in B_R(x_0)$$

for some $R > 0$ and

$$\sup_{B_{\frac{R}{2}}(x_0)} u(x,t_2) > 0.$$

Furthermore, without loss of generality we may assume that

$$s = \frac{t_0 - t_2}{t_2 - t_1}$$

is sufficiently large. Hence, from Lemma 5 we conclude that

$$\fint_{B_R(x_0)} u^p(x,t_2)\, dx \ge \frac{1}{c}\left(\frac{R^p}{t_2 - t_1}\right)^{\frac{p}{p-2}}$$

and

$$\fint_{B_R(x_0)} u^p(x,t_2)\,dx \geq \frac{1}{c}\left(\frac{t_0-t_2}{t_2-t_1}\right)^{\frac{p}{p-2}}\left(\frac{R^p}{t_0-t_2}\right)^{\frac{p}{p-2}}.$$

Now we recall Lemma 3 and obtain

$$\begin{aligned}\frac{1}{c}\left(\frac{t_0-t_2}{t_2-t_1}\right)^{\frac{p}{p-2}}\left(\frac{R^p}{t_0-t_2}\right)^{\frac{p}{p-2}} &\leq \fint_{B_R(x_0)} u^p(x,t_2)\,dx \\ &\leq c\left[\left(\frac{R^p}{t_0-t_2}\right)^{\frac{p}{p-2}}+\left(\frac{t_0-t_2}{R^p}\right)^{n} u(x_0,t_0)^{\kappa}\right].\end{aligned}$$

Thus if $(t_0-t_2)/(t_2-t_1)$ is large enough, we have

$$\left(\frac{t_0-t_2}{R^p}\right)^{n} u(x_0,t_0)^{\kappa} \geq c\left(\frac{R^p}{t_0-t_2}\right)^{\frac{p}{p-2}} > 0,$$

and this contradicts the fact that $(x_0,t_0)\in\Gamma$ and $u(x_0,t_0)=0$. □

LEMMA 7. *Γ_1 is relatively open in Γ and Γ_2 is relatively closed in Γ.*

Proof. We need only to show that Γ_2 is closed. Let (x_0,t_0) be a limit point of Γ_2; then there is a sequence of points $(x_k,t_k)\in\Gamma_2$ such that

$$(x_k,t_k)\to(x_0,t_0).$$

Since $x_k\in\Gamma(0)$, we note that $x_0\in\Gamma(0)$. Considering Lemma 6 we know that $\Gamma_1\cup\Gamma_2=\Gamma$. Therefore we conclude that

$$(x_0,t_0)\in\Gamma_2. \qquad \square$$

Now we prove that the rate of the growth of Γ_1 is Hölder continuous.

THEOREM 4. *Suppose that $(x_0,t_0)\in\Gamma_1$, that is, the vertical segment does not contain any point of Γ. Here we assume t_0 is a certain positive time. Then there exist constants c, h, and α such that*

$$u(x,t)=0 \quad \textit{for } t_0-h\leq t\leq t_0 \textit{ and } |x-x_0|\leq c(t_0-t)^{\alpha}$$

and

$$u(x,t)>0 \quad \textit{for } t_0<t\leq t_0+h \textit{ and } |x-x_0|\leq c(t_0-t)^{\alpha}.$$

Proof. Let $t_1<t_0$ be fixed and $h=t_0-t_1$. From Lemma 6 we know that there exists R such that $B_R(x_0)\cap\Omega(t_1)=\emptyset$, that is,

$$u(x,t_1)=0 \quad \text{for all } x\in B_R(x_0).$$

Let $t=t_1+\delta h$, where δ is fixed later. From Lemma 5 we see that if

$$\operatorname{dist}(x_0,\Omega(t))<dR,$$

then for some $x_1\in\Omega(t)$ with $\operatorname{dist}(x_1,x_0)=dR$,

$$\fint_{B_{(1-d)R}(x_1)} u^p(x,t)\,dx > \frac{1}{c}\left[\frac{(1-d)^pR^p}{\delta h}\right]^{\frac{p}{p-2}},$$

where $d < \frac{1}{4}$ is a small number fixed later. Thus we obtain

$$\fint_{B_R(x_0)} u^p(x,t)\,dx > \frac{1}{c}\frac{(1-d)^n(1-d)^{\frac{p^2}{p-2}}}{\delta^{\frac{p}{p-2}}}\left[\frac{R^p}{h}\right]^{\frac{p}{p-2}}.$$

Again, as in the proof of Lemma 6 we have

$$\begin{aligned}\frac{1}{c}\frac{(1-d)^n(1-d)^{\frac{p^2}{p-2}}}{\delta^{\frac{p}{p-2}}}\left[\frac{R^p}{h}\right]^{\frac{p}{p-2}} &\le \fint_{B_R(x_0)} u^p(x,t)\,dx \\ &\le B\left[\left(\frac{R^p}{t_0-t}\right)^{\frac{p}{p-2}} + \left(\frac{t_0-t}{R^p}\right)^n u(x_0,t_0)^\kappa\right] \\ &\le B\left[\left(\frac{R^p}{(1-\delta)h}\right)^{\frac{p}{p-2}} + \left(\frac{(1-\delta)h}{R^p}\right)^n u(x_0,t_0)^\kappa\right],\end{aligned}$$

where B is the constant appearing in (7). Hence we obtain

$$\left[\frac{1}{c}\frac{(1-d)^n(1-d)^{\frac{p^2}{p-2}}}{\delta^{\frac{p}{p-2}}} - \frac{B}{(1-\delta)^{\frac{p}{p-2}}}\right]\left(\frac{R^p}{h}\right)^{\frac{p}{p-2}}$$

$$\le B\left(\frac{(1-\delta)h}{R^p}\right)^n u(x_0,t_0)^\kappa.$$

On the other hand, if δ is small and d is near 0, then

$$\frac{1}{c}\frac{(1-d)^n(1-d)^{\frac{p^2}{p-2}}}{\delta^{\frac{p}{p-2}}} - \frac{B}{(1-\delta)^{\frac{p}{p-2}}} > 0,$$

and this contradicts the fact that

$$u(x_0,t_0) = 0.$$

Thus we have

$$\operatorname{dist}(x_0,\Gamma(t)) \ge dR.$$

We set $d = (1-\delta)^\alpha$. Hence we have

$$\operatorname{dist}(x_0,\Gamma(t_0-(1-\delta)h)) \ge (1-\delta)^\alpha R.$$

Repeating the above process with $t_1 = t$, we obtain

$$\operatorname{dist}\left(x_0,\Gamma(t_0-(1-\delta)^2h)\right) \ge (1-\delta)^{2\alpha}R.$$

In a similar way, we can iterate the above process for all $k \ge 1$ and conclude that

$$\operatorname{dist}\left(x_0,\Gamma(t_0-(1-\delta)^kh)\right) \ge (1-\delta)^{k\alpha}R$$

for all k. Varying h we conclude that

$$\operatorname{dist}(x_0,\Gamma(t)) \ge \left(\frac{t_0-t}{h}\right)^\alpha R,$$

and this completes the proof for the first claim. The second claim can be proved in the same way. □

Considering Theorem 1, we find that if

$$u_0(x) \geq [\mathrm{dist}(x)]^{\frac{\gamma}{p-2}}, \quad \gamma < p$$

for all $x \in \Omega(0)$, then $\Gamma = \Gamma_1$. Moreover, Theorem 4 implies that the interface is given by a function

$$t = S(x),$$

and if $S(x_0) \geq \eta_0$ for some fixed $\eta_0 > 0$, then

$$|S(x) - S(x_0)| \leq c|x - x_0|^{\frac{1}{\alpha}}$$

for some c depending on η_0. Hence $\Omega(t+h)$ contains a (ch^α) neighborhood of $\Omega(t)$ for $0 < h < 1$. Now we find a bound for the velocity of the interface.

THEOREM 5. *For any $\eta_0 > 0$ there exists a positive constant c depending only on p, n, η_0 such that for any $t > \eta_0, 0 < h < 1$,*

$$\Gamma(t+h) \text{ is contained in the } \left(ch^{\frac{1}{p}}\right) \text{ neighborhood of } \Gamma(t).$$

Proof. Suppose that $u(x_0, t_0) = 0$ and $\mathrm{dist}\,(x_0, \Gamma(t_0)) = a$. Let

$$v(x,t) = \lambda \left([\alpha^p (t - t_0) + \alpha(|x - x_0| - b)]^+ \right)^q,$$

where $q = \frac{p-1}{p-2}$ and α and b are decided later. Observe that

$$(q-1)(p-1) = q.$$

With a direct computation we obtain

$$\begin{aligned} & v_t - \mathrm{div}\left(|\nabla v|^{p-2}\nabla v\right) \\ &= \lambda q \alpha^p \left([\alpha^p (t - t_0) + \alpha(|x - x_0| - b)]^+ \right)^{\frac{1}{p-2}} \\ &\quad \cdot \left(1 - \lambda^{p-2} q^{p-1} - \lambda^{p-2} q^{p-2} \frac{(n-1)}{\alpha |x - x_0|} [\alpha^p (t - t_0) + \alpha(|x - x_0| - b)]^+ \right). \end{aligned}$$

Thus, if

$$1 - \lambda^{p-2} q^{p-1} - \lambda^{p-2} q^{p-1} \frac{(n-1)}{|x - x_0|} \left[\alpha^{p-1}(t - t_0) + (|x - x_0| - b)\right]^+ \geq 0,$$

that is,

$$\lambda^{p-2} q^{p-1} + \lambda^{p-2} q^{p-1} \frac{(n-1)}{|x - x_0|} \left[\alpha^{p-1}(t - t_0) + (|x - x_0| - b)\right]^+ \leq 1, \tag{14}$$

then v is a supersolution.

Now we take α and b such that

$$\alpha^q (a - b)^q = \frac{M}{\lambda},$$

where

$$M = \sup u.$$

With this choice of α, λ, and b we find that $v(x,t) \geq u(x,t)$ for all $x, |x - x_0| = a$, and $t_0 \leq t \leq t_1$, where t_1 is a certain fixed time. By the usual comparison principle we see that

$$u(x,t) \leq v(x,t)$$

for all $x \in B_a(x_0)$ and $t_0 \leq t \leq t_1$. Note that the interface of $v(x, t_0 + h)$ is decided by

$$|x - x_0| = b - \alpha^{p-1} h.$$

We take α satisfying

$$\frac{1}{\alpha}\left(\frac{M}{\lambda}\right)^{\frac{1}{q}} = \alpha^{p-1} h,$$

and hence

$$\alpha = \left(\frac{1}{h}\right)^{\frac{1}{p}} \left(\frac{M}{\lambda}\right)^{\frac{1}{pq}}.$$

Therefore, the interface of $v(x, t_0 + h)$ is

$$\begin{aligned} |x - x_0| = b - \alpha^{p-1} h &= a - \frac{1}{\alpha}\left(\frac{M}{\lambda}\right)^{\frac{1}{q}} - \alpha^{p-1} h \\ &= a - 2\left(\frac{M}{\lambda}\right)^{\frac{p-1}{pq}} h^{\frac{1}{p}}. \end{aligned}$$

Taking λ so small that

$$\lambda^{p-2} q^{p-1} + \lambda^{p-2} q^{p-1} \frac{(n-1)}{|x - x_0|} \left[\alpha^{p-1}(t - t_0) + (|x - x_0| - b)\right]^+ \leq 1,$$

we see that $\Gamma(t_0 + h)$ is contained in the $ch^{\frac{1}{p}}$ neighborhood of $\Gamma(t_0)$. □

4. Lipschitz continuity of the interface after a large time. Following the argument in [7], Lipschitz regularity of the interface after a large time is proved. As in the case of porous medium equations we have a monotonicity property after a large time based on the Alexandrov reflection principle. In this section we assume that the support of the initial datum u_0 is contained in $B_{R_0}(0)$.

LEMMA 8. *Let $x_0, x_1 \in R^n, |x_0|, |x_1| > R_0$, and*

$$\cos\langle x_1 - x_0, x_0\rangle \geq \frac{R_0}{|x_0|}.$$

Then for every $t > 0$ we have

$$u(x_1, t) \leq u(x_0, t).$$

Since Lemma 8 follows from a comparison principle and the reflection principle as in the case of porous medium equations, we refer to [7] for the proof.

Considering the asymptotic behaviour of u, there exists a large time t_0 such that

$$B_{R_0}(0) \subset \Omega(t)$$

for all $t > t_0$. Hence $(x,t) \in \Gamma$ implies that $|x| > R_0$ for all $t > t_0$. Hence, by virtue of Lemma 8 we obtain the following corollary. Indeed the same corollary for porous medium equations appears in [7].

COROLLARY 1. *There exists t_0 such that the interface $\Gamma(t)$ is representable in polar coordinates as follows*

$$r = f(\theta, t), \qquad f(\theta, t) \geq R_0$$

for all $t > t_0$, and f is Lipschitz in θ.

Proof. It follows from Lemma 8 that for every $(\bar{x}, \bar{t}) \in \Gamma, |\bar{x}| > R > R_0, \bar{t} > t_0$, we have $u(x,t) = 0$ for every x in a cone

$$K_\varepsilon = \left\{ x : |x - \bar{x}| < \varepsilon \text{ and } \cos\langle x - \bar{x}, \bar{x}\rangle \geq \frac{R}{|\bar{x}|} \right\}$$

and $u(x,t) > 0$ for every x in a cone

$$K_\varepsilon = \left\{ x : |x - \bar{x}| < \varepsilon \text{ and } \cos\langle x - \bar{x}, \bar{x}\rangle \leq -\frac{(1+\varepsilon)R}{|\bar{x}|} \right\}$$

for some small ε. This implies that f is Lipschitz in θ. □

Considering scaling and asymptotic behaviour of u we have an estimate of $\sup|\nabla u|$ in terms of u after a large time.

LEMMA 9. *For every $\varepsilon > 0$ there exists $T_1 = T_1(\varepsilon, R_0, u_0)$ such that*

$$|\nabla u(x,t)| \leq \varepsilon u(x,t) \tag{15}$$

for all $|x| \leq R_0$ and $t \geq T_1$.

Proof. We define a family of functions

$$u_k(x,t) = k^n u(kx, k^\kappa t)$$

with parameter k, where $\kappa = n(p-2) + p$. We note that u_k are again solutions of (1). We know that u_k converge to a fundamental solution $\bar{u}$ uniformly with respect to $t \geq \tau$ for every τ (see Theorem 3 in [16]). Therefore there exists k_0 such that

$$c > 2\bar{u}(x,t) \geq u_k(x,t) \geq \frac{1}{2}\bar{u}(x,t) > \frac{1}{c} > 0$$

for $|x| \leq 2R_0, \frac{1}{2} \leq t \leq 2$, $k \geq k_0$, and for some c depending only on u_0, n, and p. Since u_k are uniformly bounded in $R^n \times (\frac{1}{2}, 2)$, we obtain

$$|\nabla u_k| \leq c_1$$

for $t = 1$, $|x| \leq R_0$, and some c_1, and this implies

$$|\nabla u(y, k^\kappa)| \leq c k^{-(n+1)}$$

for all $|y| < R_0$. Note that $u(x, k^\kappa) \geq ck^{-n}$ for all $|x| \leq R_0$. Setting $t = k^\kappa$ we conclude that

$$|\nabla u(x,t)| \leq ct^{-\frac{1}{\kappa}} u(x,t)$$

if $|x| \leq R_0, t = k^\kappa$, and $k \geq k_0$. This completes the proof. □

Finally, we show that the interface is representable in a polar coordinate with a Lipschitz function. In fact, from Corollary 1 we need only to show that f in Corollary 1 is Lipschitz in t.

THEOREM 6. *There exists t_1 such that $\Gamma(t), t > t_1$ can be represented by*

$$r = f(\theta, t),$$

where f is locally Lipschitz continuous in θ and t.

Proof. Considering Corollary 1, we know that f is Lipschitz in θ and thus we need only to prove that f is Lipschitz in t. We define a family of solutions to (1) in $R^n \times (t_1, \infty)$:

$$u_\varepsilon(x,t) = \frac{1}{(1+\varepsilon)^{\frac{p-1}{p-2}}} u((1+\varepsilon)x, (1+\varepsilon)t + t_1)$$

for $\varepsilon > 0$. Here t_1 is a large time decided later. We want to show that for every $\varepsilon \in (0,1)$ and $x \in R^n$,

$$u_\varepsilon(x,0) \leq u(x,t_1).$$

To do this we write the difference $u_\varepsilon(x,0) - u(x,t_1)$ as

$$u_\varepsilon(x,0) - u(x,t_1) = \left[\frac{1}{(1+\varepsilon)^{\frac{p-1}{p-2}}} - 1\right] u((1+\varepsilon)x, t_1) + u((1+\varepsilon)x, t_1) - u(x,t_1).$$

If $|x| > R_0$, then from Lemma 8,

$$u((1+\varepsilon)x, t_1) - u(x,t_1) \leq 0$$

and

$$u_\varepsilon(x,0) - u(x,t_1) \leq \left[\frac{1}{(1+\varepsilon)^{\frac{p-1}{p-2}}} - 1\right] u((1+\varepsilon)x, t_1) \leq 0.$$

Now we consider the case $|x| \leq R_0$. From Lemma 9, for t_1 large enough we have

$$u((1+\varepsilon)x, t_1) - u(x,t_1) \leq \varepsilon |x| |\nabla u(\xi x, t_1)| \leq \varepsilon R_0 \frac{1}{cR_0} u(x,t_1) \leq \frac{\varepsilon}{c} u(x,t_1)$$

with some large constant c, where $\xi \in (1, 1+\varepsilon)$. Hence we have

$$\frac{u_\varepsilon(x,0) - u(x,t_1)}{\varepsilon} \leq 0$$

for all $\varepsilon \in (0,1)$, and differentiating $u_\varepsilon(x,t_1)$ with respect to ε we have

$$-\frac{p-1}{p-2} u(x, t+t_1) + x \cdot \nabla u(x, t+t_1) + t u_t(x, t+t_1) \leq 0.$$

Replacing t by $t+t_1$ we obtain

$$(t-t_1)u_t(x,t) \le \frac{p-1}{p-2}u(x,t) + x\cdot\nabla u(x,t). \tag{16}$$

Hence u_t is bounded. Setting $t-t_1=h$ for some fixed h, (16) can be written in the form

$$\frac{d}{dt}\left[e^{-\frac{(p-1)t}{(p-2)h}}u(r_0e^{\frac{t-t_2}{h}},\theta,t)\right] \le 0$$

for $t-h<t_2<t$ and θ fixed. Therefore, if $u(r_0,\theta,t_2)=0$, then

$$u(r_0e^{\frac{t-t_2}{h}},\theta,t)=0 \quad \text{for } t>t_2.$$

This gives

$$f(\theta,t) \le f(\theta,t_2)e^{\frac{t-t_2}{h}},$$

and we get for $t-h<t_2<t$,

$$f(\theta,t)-f(\theta,t_2) \le c_3 f(\theta,t_2)(t-t_2),$$

where $c_3=c_3(\frac{1}{h})$. □

5. Global Lipschitz regularity. Employing the method of [7], we show that the interface is Lipschitz if the initial datum satisfies nondegeneracy conditions. Define $v=u^{(p-2)/(p-1)}$; then one can formally prove that v satisfies

$$v_t = c_1 v\operatorname{div}\left(|\nabla v|^{p-2}\nabla v\right) + c_2|\nabla v|^p \tag{17}$$

with initial data $v(x,0)=u_0^{(p-2)/(p-1)}$, where $c_1=((p-1)/(p-2))^{p-2}$ and $c_2=((p-1)/(p-2))^{p-1}$.

Set $\Omega=\Omega(0)$. Let $v_0=(u_0)^{(p-2)/(p-1)}$ We assume the following:

(i) v_0 is integrable and positive in a C^1 domain $\Omega\subset B_R(0)$.

(ii) $v_0\in W^{1,1}(S)$ in a certain strip $S\subset\Omega$ along the boundary $\partial\Omega$ and there exist constants K_1 and K_2 such that

$$K_1 \le |\nabla v_0| \le K_2$$

in S.

(iii) There exists $a>0$ such that

$$v_0 > a \quad \text{in} \quad \Omega\setminus S.$$

(iv) There exists a constant K_0 such that

$$\nabla^2 v_0 \ge -K_0 I$$

in the sense of distributions.

LEMMA 10. *There exist constants $A,B>0$ such that*

$$\frac{A-p}{p-1}v(x,t) + x\cdot\nabla v(x,t) + (At+B)v_t \ge 0. \tag{18}$$

Proof. We consider a family of solutions

$$v^\varepsilon(x,t) = \frac{(1+A\varepsilon)^{\frac{1}{p-1}}}{(1+\varepsilon)^{\frac{p}{p-1}}} v((1+\varepsilon)x, (1+A\varepsilon)t + B\varepsilon)). \tag{19}$$

We show that

$$v^\varepsilon(x,t) \geq v(x,t)$$

for small ε, and then differentiating v^ε with respect to ε we have

$$\frac{A-p}{p-1} v(x,t) + x \cdot \nabla v + (At+B)v_t \geq 0.$$

We approximate v_0 by

$$v_0^\delta = v_0 * \rho_\delta(x) + \delta^\alpha,$$

where $\rho_\delta(x)$ is a convolution kernel and α is decided later. Suppose that $v^\delta(x,t) = \left(u^\delta(x,t)\right)^{(p-2)/(p-1)}$ is the solution to

$$v_t^\delta = c_1 v^\delta \mathrm{div}\left(|\nabla v^\delta|^{p-2}\nabla v^\delta\right) + c_2 |\nabla v^\delta|^p$$

with $v^\delta(x,0) = v_0^\delta(x)$. We note that $v^\delta \geq \delta^\alpha > 0$ and $v^\delta \in C^\infty$. If there is no confusion, we omit δ in various expressions.

If δ is sufficiently small, then

$$v(x,0) \geq \frac{a}{2} \quad \text{in } \Omega_1 = \Omega \setminus S$$

and

$$|\nabla v(x,0)| \geq \frac{K_1}{4} \quad \text{in } S.$$

In fact, this inequality is also true in a neighborhood $U_{c\delta}$ of $\partial\Omega$ of the form

$$U_{c\delta} = \{x \in R^n; \mathrm{dist}(x, \partial\Omega) < c\delta\}$$

for a constant $c \in (0, \frac{1}{2})$.

Now we consider several different regions.

(i) First we consider the region where $|\nabla v_0| > K_1/4$ (in particular, $S \cup U_{c\delta}$). We have

$$\begin{aligned} I_\varepsilon &\equiv \frac{1}{\varepsilon}\left(v^\varepsilon(x,0) - v(x,0)\right) \\ &\geq \frac{1}{\varepsilon}\left[\left(1 + \frac{\varepsilon}{2}\frac{A-p}{p-1}\right) v((1+\varepsilon)x, B\varepsilon) - v(x,0)\right] \end{aligned}$$

if ε is small. So from the mean value theorem,

$$\begin{aligned} I_\varepsilon \geq{}& \frac{1}{2}\frac{A-p}{p-1} v((1+\varepsilon)x, B\varepsilon) \\ &+ B v_t((1+\varepsilon)x, \theta\varepsilon) + x \cdot \nabla v(\xi, 0), \end{aligned}$$

where $\theta \in (0, B)$ and ξ lies in the segment $\overline{x, (1+\varepsilon)x}$. If ε is small enough, then

$$\begin{aligned} &v((1+\varepsilon)x, B\varepsilon) \cong v(x,0), \\ &v_t((1+\varepsilon)x, \theta\varepsilon) \cong v_t(x,0), \\ &\nabla v(\xi, 0) \cong \nabla v(x,0). \end{aligned}$$

Therefore, there exists $C > 0$ depending only on n, p, and v_0^δ such that

$$\begin{aligned} I_\varepsilon \geq \frac{1}{2}\frac{A-p}{p-1} v(x,0) + B v_t(x,0) \\ + x \cdot \nabla v(x,0) - c\varepsilon \end{aligned}$$

for some c. Using the equation

$$v_t = c_1 v \Delta_p v + c_2 |\nabla v|^p,$$

we get

$$\begin{aligned} I_\varepsilon \geq \left(\frac{1}{2}\frac{A-p}{p-1} + B c_1 \Delta_p v(x,0) \right) v(x,0) \\ + B c_2 |\nabla v(x,0)|^p + x \cdot \nabla v(x,0) - c\varepsilon. \end{aligned}$$

Since $|x| \leq R + \delta$ and $\Delta v_0 \geq -nK_0$, we have

$$\begin{aligned} I_\varepsilon \geq \left(\frac{1}{2}\frac{A-p}{p-1} - B c_1 n K_0 \right) v(x,0) \\ + |\nabla v(x,0)| \left(B c_2 |\nabla v|^{p-1} - R - \delta \right) - c\dot{\varepsilon}. \end{aligned}$$

If we choose A and B such that

$$B > \frac{4R}{c_2 K_1^{p-1}} \quad \text{and} \quad \frac{1}{2}\frac{A-p}{p-1} - B c_1 n K_0 \geq 0,$$

then

$$I_\varepsilon \geq \frac{K_1}{4} \left(\frac{B c_2 K_1^{p-1}}{4} - R - \delta \right) - c\varepsilon > 0$$

for small ε and δ.

(ii) Next, we consider the region $\Omega_1 = \Omega \setminus S$.

We only need to consider those points where

$$|\nabla v_0(x)| \leq \frac{K_1}{4}.$$

So in this case we see that

$$I_\varepsilon \geq \left(\frac{1}{2}\frac{A-p}{p-1} - B c_1 n K_0 \right) \frac{a}{2} - \frac{R K_1}{4} - c\varepsilon,$$

and if $\frac{1}{2}\frac{A-p}{p-1} - B c_1 n K_0 \geq R K_1 / a$, then

$$I_\varepsilon \geq 0$$

for small ε.

(iii) Next, we consider the region $\Omega_3 = \{x \in R^n : \operatorname{dist}(x, \Omega) \geq \delta\}$.

In Ω_3, $v(x,0) = \delta^\alpha$. Since $v \geq \delta^\alpha$ from the maximum principle, we have

$$I_\varepsilon \geq 0.$$

(iv) Finally, we consider the region $\Omega_4 = \{x \in R^n; c\delta \leq \operatorname{dist}(x, \Omega) \leq \delta\}$ with $0 < c < 1$.

In this case we select a particular cutoff function $\{\rho_\delta\}$ satisfying

$$\begin{aligned}
&\rho_\delta(x) = 0 \quad \text{if} \quad |x| \geq \delta, \\
&\rho_\delta(x) = \rho_\delta(0) \quad \text{if} \quad |x| \leq \delta - \delta^{1+\gamma} \quad \text{for certain } 0 < \gamma < 1, \\
&0 \leq \rho_\delta(x) \leq \rho_\delta(0), \quad \rho_\delta \in C^\infty.
\end{aligned}$$

Now suppose $\operatorname{dist}(x, \Omega) \in (\delta - \delta^{1+\gamma}, \delta)$; then

$$\begin{aligned}
\left|\nabla v_0^\delta(x)\right| &\leq \int_{B_\delta \cap \Omega} |\nabla v(y)| \rho_\delta(x-y) dy \\
&\leq cK_2 \int_{B_\delta \cap \Omega} \rho_\delta(x-y) dy.
\end{aligned}$$

Now we observe that

$$\int_{B_\delta \cap \Omega} \rho_\delta(x-y) dy \leq c\delta^{\gamma \frac{n+1}{2}}$$

and hence

$$\left|\nabla v_0^\delta(x)\right| \leq cK_2 \delta^{\gamma \frac{n+1}{2}}.$$

Thus

$$I_\varepsilon \geq \left(\frac{1}{2}\frac{A-p}{p-1} - Bc_1 nK_0\right)\delta^\alpha - (R+2\delta)K_2\delta^{\gamma\frac{n+1}{2}} - c\varepsilon.$$

In particular, if $0 < \frac{2\alpha}{n+1} < \gamma < 1$ and ε is small, then

$$I_\varepsilon \geq 0.$$

Finally, we consider those points x such that

$$c\delta \leq \operatorname{dist}(x, \Omega) \leq \delta - \delta^{\gamma+1}.$$

Recall from assumption (iv) that

$$\nabla^2 v_0 \geq -K_0 I$$

in the sense of distributions. We know that

$$\begin{aligned}
I_\varepsilon \geq &\left[\frac{1}{2}\frac{A-p}{p-1} + Bc_1 |\nabla v|^{p-2} a_{ij}(\nabla v)(v)_{x_i x_j}\right] v_\delta(x,0) \\
&+ Bc_2 |\nabla v(x,0)|^p - |x| |\nabla v_\delta(x,0)| - c\varepsilon,
\end{aligned}$$

where

$$a_{ij}(\nabla v) = \delta_{ij} + (p-2)\frac{(v)_{x_j}(v)_{x_j}}{|\nabla v(x,0)|^2}.$$

Since $v_0 \in W^{1,1}(S)$, we have that

$$\begin{aligned}|\nabla v(x,0)| &= \left|\int_{B_\delta \cap \Omega} \nabla v_0(y)\rho_\delta(x-y)dy\right| \\ &\geq K_1 \int_{B_\delta \cap \Omega} \rho_\delta(x-y)dy \\ &\geq cK_1\delta^{\gamma\frac{n+1}{2}}.\end{aligned}$$

Also, we get from assumption (iv) that

$$\Delta v = \int_{B_\delta \cap \Omega} \Delta v_0(y)\rho_\delta(x-y)dy - \int_{\partial(B_\delta \cap \Omega)} (\nabla v_0 \cdot \nu)\,\rho_\delta(x-y)d\sigma_y$$

and

$$\begin{aligned}|\Delta v| &\geq K_1 \int_{B_\delta \cap \partial\Omega} \rho_\delta(x-y)d\sigma_y - nK_0 \int_{B_\delta \cap \Omega} \rho_\delta(x-y)dy \\ &\geq K_1\delta^{-1+\frac{\gamma(n-1)}{2}} - nK_0\delta^{\frac{\gamma(n+1)}{2}}.\end{aligned}$$

Observe that at $x \in \partial\Omega$,

$$\nabla v_0 = c\nu.$$

Hence we obtain

$$\begin{aligned}&(p-2)\frac{v_{x_j}v_{x_j}}{|\nabla v|^2}\int_{B_\delta \cap \Omega} v_0(y)\left(\rho_\delta(x-y)\right)_{x_ix_j}dy \\ &= (p-2)\frac{v_{x_j}v_{x_j}}{|\nabla v|^2}\int_{B_\delta \cap \Omega} (v_0)_{x_ix_j}(y)\rho_\delta(x-y)dy \\ &-(p-2)\frac{v_{x_j}v_{x_j}}{|\nabla v|^2}\int_{B_\delta \cap \partial\Omega} (v_0)_{x_i}\nu_j\rho_\delta(x-y)d\sigma_y \\ &\geq -cK_0\delta^{\gamma\frac{n+1}{2}}\end{aligned}$$

for some c. Therefore, combining all these together, we obtain

$$\begin{aligned}&Bc_1|\nabla v|^{p-2}a_{ij}(\nabla v)(v)_{x_ix_j}v(x,0) - (R+2\delta)|\nabla v| \\ &\geq Bc_1c\left[K_1\delta^{\gamma\frac{n+1}{2}}\right]^{p-2}\left(K_1\delta^{-1+\frac{\gamma}{2}(n-1)} - cK_0\delta^{\gamma\frac{n+1}{2}}\right)\delta^\alpha \\ &-(R+2\delta)\frac{K_1}{5} - c\varepsilon.\end{aligned}$$

Hence, for sufficiently small δ we have

$$I_\varepsilon \geq Bc_1cK_1^{p-1}\delta^{-1+\alpha+\gamma\left(\frac{(n+1)(p-2)}{2}+\frac{n-1}{2}\right)} + (R+2\delta)\frac{K_1}{5} - c\varepsilon.$$

Therefore, if we choose $0 < \alpha < 1$ and γ so that

$$\frac{2\alpha}{1+n} < \gamma < \frac{1-\alpha}{\frac{(n+1)p}{2} - \frac{n+3}{2}},$$

then for sufficiently small ε compared to δ we get

$$I_\varepsilon \geq 0. \qquad \square$$

LEMMA 11. *Under the assumptions above, the function* $v(x(s),t(s))e^{\frac{A-p}{p-1}}$ *is non-decreasing along the curves*

$$x(s) = x_0 e^s, \quad t(s) = \frac{1}{A}\left[(At_0 + B)\, e^{As} - B\right], \ s \geq 0.$$

Proof. Along these curves,

$$x'(s) = x(s) \ \text{ and } \ t'(s) = At(s) + B.$$

Hence, by (18) we have

$$\frac{d}{ds}\left[v(x(s),t(s))e^{\frac{A-p}{p-1}s}\right] = \left[\frac{A-p}{p-1}v + x\cdot\nabla v + (At+B)v\right] e^{\frac{A-p}{p-1}s} \geq 0. \qquad \square$$

COROLLARY 2.1. If $r = f(\theta, t_1)$ is the equation of the free boundary for $t_1 > t_0$, then

$$f(\theta,t_1) \geq f(\theta,t_0)\left(\frac{At_1+B}{At_0+B}\right)^{\frac{1}{A}}.$$

Proof. The curves $(x(s),t(s))$ above can be written as

$$x(s) = x_0 e^s = x_0\left(\frac{At_1+B}{At_0+B}\right)^{\frac{1}{A}}. \tag{20}$$

So, if $\theta = x_0/|x_0|$ and $r_0 = |x_0|$, we have that $v(r_0,\theta,t_0) > 0$ implies

$$v\left(r_0\left(\frac{At_1+B}{At_0+B}\right)^{\frac{1}{A}}, \theta, t_0\right) > 0.$$

Therefore we obtain

$$f(\theta,t) \geq r_0\left(\frac{At_1+B}{At_0+B}\right)^{\frac{1}{A}} \geq f(\theta,t_0)\left(\frac{At_1+B}{At_0+B}\right)^{\frac{1}{A}}. \qquad \square$$

The following theorem implies that the free boundary is Lipschitz.

THEOREM 7. *For every point* $(\bar{x},\bar{t})$ *there exist positive constants* $A,B,C > 0$ *depending only on* $v_0, n, p, \bar{t}$ *and* $R_1 = \sup\{\mathrm{dist}(\bar{x},y); y \in \Omega\}$ *such that*

$$v(x,t) \geq v(\bar{x},\bar{t})e^{-C(t-\bar{t})} \tag{21}$$

for every (x,t) *satisfying* $\bar{t} < t < \bar{t} + \varepsilon$ *for some* $\varepsilon = \varepsilon(A,B)$ *and*

$$\frac{|x-\bar{x}|}{|t-\bar{t}|} \leq \frac{R_1}{A\bar{t}+B}.$$

Proof. Let us take A, B, and $C = \frac{A-p}{p-1}$ as in Lemma 10 with $R = 3R_1$. The inequality will be true if every (x,t) in this conical region around $(\bar{x},\bar{t})$ can be decided in the form $(x(t),t)$ given by (20) with a particular origin of coordinates such that Ω is contained in the ball with center at that origin and radius R. The origin x_0 corresponding to (x,t) will be given by

$$x - x_0 = (\bar{x} - x_0)\left(\frac{At+B}{A\bar{t}+B}\right)^{\frac{1}{A}}.$$

Therefore

$$\begin{aligned}|\bar{x} - x_0|\left(\frac{At+B}{A\bar{t}+B}\right)^{\frac{1}{A}} &\le |x-\bar{x}| + |\bar{x}-x_0| \\ &\le R_1\frac{t-\bar{t}}{A\bar{t}+B} + |\bar{x}-x_0|.\end{aligned}$$

So we have

$$|\bar{x} - x_0|\left[\left(\frac{At+B}{A\bar{t}+B}\right)^{\frac{1}{A}} - 1\right] \le R_1\frac{t-\bar{t}}{A\bar{t}+B}.$$

Writing $(\frac{At+B}{A\bar{t}+B})^{1/A}$ as a power of $t-\bar{t}$ and using the fact that ε is small, we get

$$\frac{1}{2}|\bar{x} - x_0|\frac{t-\bar{t}}{A\bar{t}+B} \le R_1\frac{t-\bar{t}}{A\bar{t}+B},$$

that is, $|\bar{x} - x_0| \le 2R_1$. This gives

$$|x_0 - y| \le 3R_1 = R$$

for every $y \in \Omega$. Therefore inequality (21) holds. □

Acknowledgment. The authors express deep thanks to Professor E. DiBenedetto for his comments and suggestions. In particular, Theorem 1 in §2 was suggested by Professor E. DiBenedetto.

REFERENCES

[1] D. G. Aronson, *Regular properties of flows through porous media: The interface*, Arch. Rational Mech. Anal., 37 (1970), pp. 1–10.

[2] D. G. Aronson and Ph. Benilan, *Regularité des solutions de l'equation des milieux poreus dans R^n*, C. R. Acad. Sci. Paris, Sér. A-B, 288 (1979), pp. 103–105.

[3] Ph. Benilan and M. G. Crandall, *Regularizing effects of homogeneous evolution equations*, MRC Tech. report 2076, University of Wisconsin, Madison, Wisconsin, 1980.

[4] Ph. Benilan, M. G. Crandall, and M. Pierre, *Solutions of porous medium equation in R^n under optimal conditions on initial values*, Indiana Univ. Math. J., 33 (1984), pp. 51–87.

[5] L. A. Caffarelli and A. Friedman, *Regularity of the free boundary of a gas flow in an n-dimensional porous medium*, Indiana Univ. Math. J., 29 (1980), pp. 361–389.

[6] L. A. Caffarelli and N. Wolanski, *$C^{1,\alpha}$ regularity of the free boundary for the n-dimensional porous medium equation*, Comm. Pure Appl. Math., 43 (1990), pp. 885–902.

[7] L. A. Caffarelli, J. Vazquez, and N. Wolanski, *Lipschitz continuity of solutions and interface of the n-dimensional porous medium equation*, Indiana Univ. Math. J., 36 (1987), pp. 373–401.

[8] H. J. CHOE, *Hölder continuity for solutions of certain degenerate parabolic system*, Nonlinear Anal., 18 (1992), pp. 235–243.

[9] E. DIBENEDETTO, *Intrinsic Harnack type inequalities for solutions of certain degenerate parabolic equations*, Arch. Rational Mech. Anal., 100 (1988), pp. 129–147.

[10] ———, *On the local behaviour of solutions of degenerate parabolic equations with measurable coefficients*, Ann. Scuola. Norm Sup. Pisa Cl. Sci., (4), 13 (1986), pp. 487–535.

[11] E. DIBENEDETTO AND M. HERRERO, *Non-negative solutions of the evolution p-Laplacian equations, initial traces and Cauchy problems when* $1 < p < 2$, Arch. Rational Mech. Anal., 111 (1990), pp. 225–290.

[12] ———, *On the Cauchy problem and initial traces for a degenerate parabolic equation*, Trans. Amer. Math. Soc., 314 (1989), pp. 187–224.

[13] E. DIBENEDETTO AND A. FRIEDMAN, *Regularity of solutions of nonlinear degenerate parabolic systems*, J. Reine Angew. Math., 349 (1984), pp. 83–128.

[14] ———, *Hölder estimates for nonlinear degenerate parabolic system*, J. Reine Angew. Math., 357 (1985), pp. 1–22.

[15] J. R. ESTEBAN AND J. VAZQUEZ, *Régularité des solutions positives de l'équation parabolique p-laplacienne*, C. R. Acad. Sci. Paris Ser. I Math., 310 (1990), pp. 105–110.

[16] Z. JUNNING, *The asymptotic behaviour of solutions of a quasilinear degenerate parabolic equation*, J. Differential Equations, 103 (1993), pp. 33–52.

[17] B. F. KNERR, *The porous medium equation in one dimension*, Trans. Amer. Math. Soc., 234 (1977), pp. 381–415.

[18] O. A. LADYZENSKAYA, *New equations for the description of motion of viscous incompressible fluids and solvability in the large of boundary values problems for them*, Proc. Steklov Inst. Math., 102 (1967), pp. 95-118.

[19] J. LEWIS, *Regularity of derivatives of solutions to certain degenerate elliptic equations*, Indiana Univ. Math. J., 32 (1983), pp. 849–858.

[20] G. LIEBERMAN, *Initial regularity for solutions of degenerate parabolic equations*, Nonlinear Anal., 14 (1990), pp. 525–536.

[21] L. K. MARTINSON AND K. B. PAPLOV, *The effect of magnetic plasticity in non-Newtonian fluids*, Magnit. Gidrodinamika, 2 (1970), pp. 50–58.

[22] M. WIEGNER, *On* C^α*-regularity of the gradient of solutions of degenerate parabolic systems*, Ann. Mat. Pura Appl., (4), 145 (1986), pp. 385–405.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 820–849, July 1995

003

ENERGY ESTIMATES RELATING DIFFERENT LINEAR ELASTIC MODELS OF A THIN CYLINDRICAL SHELL II: THE CASE OF FREE BOUNDARY*

JYRKI PIILA† AND JUHANI PITKÄRANTA†

Abstract. Four different linear models describing the elastic deformation of a thin cylindrical shell are analyzed under a given smooth normal pressure distribution. Either a bending-dominated deformation state or a "soft" membrane-dominated deformation state is assumed, and the models to be considered are (1) the standard three-dimensional model, (2) a shell model of Reissner–Mindlin type, (3) the classical shell model of Koiter, Sanders, and Novozhilov, and (4) the asymptotic shell model. Energy estimates relating the models are derived.

Key words. linear elasticity, energy estimates

AMS subject classifications. 73C02, 73C20

1. Introduction.

In [PP] we derived energy estimates relating different linear elastic models of a cylindrical shell occupying the region

$$\Omega = \left\{ (x_1, x_2, x_3) \in R^3 \mid l_1 < x_1 < l_2,\ 1 - \frac{t}{2} < \sqrt{x_2^2 + x_3^2} < 1 + \frac{t}{2} \right\},$$

where $l_2 - l_1 = 1$ and $t << 1$. Here we study the same shell geometry with $l_1 = -1$ and $l_2 = 1$, but now we do not assume any kinematical constraints on the boundary of the shell. This makes the deformation state of the shell very different compared to the case studied in [PP], where at least one end of the cylinder was set clamped. The difference arises because in a cylinder with both ends free, inextensional deformations are kinematically possible. For this reason, the deformation state is either *bending dominated* or, as we name it, *soft membrane dominated* (in distinction to the "hard" membrane-dominated case studied in [PP]), depending on the shape of the load. The present paper is a composition of [P1] and [P2], where the two cases were treated separately. As in [PP], we assume that the shell is loaded by a smoothly varying normal pressure distribution acting on the outer surface

$$\Gamma_+ = \left\{ (x_1, x_2, x_3) \in \overline{\Omega} \mid -1 < x_1 < 1, \sqrt{x_2^2 + x_3^2} = 1 + \frac{t}{2} \right\}.$$

Furthermore, to prevent rigid displacements, we assume certain orthogonality constraints on the load.

We consider five different linear elastic models of the shell problem as described above. These are as follows:

(1) the standard three-dimensional elastic model (with isotropic material),
(2) a dimensionally reduced shell model of Reissner–Mindlin type,
(3) a shell model of Kirchhoff type,
(4) the asymptotic inextensional shell theory in the bending-dominated case, and

*Received by the editors April 17, 1991; accepted for publication (in revised form) December 3, 1993.

†Helsinki University of Technology, Institute of Mathematics, Otakaari 1, SF-02150 Espoo, Finland.

(5) the asymptotic membrane theory in the soft membrane case.

Model (3) is the classical model of Koiter, Sanders, and Novozhilov (see [K], [S], [N]), and model (2) is a variation of the classical model, where the Kirchhoff–Love constraints of vanishing transverse shear strains are not imposed. Model (2) is often used (explicitly or implicitly) in finite element computations. The inextensional theory may be viewed as the analog of the Kirchhoff theory of plate bending. This is very different from the asymptotic membrane theory which is relevant in membrane-dominated situations.

The inextensional theory is obtained by assuming that the load is proportional to t^3 and then taking the limit of shell theories as $t \to 0$. Whether or not a nonzero limit deformation state is obtained in such a way is actually a test for whether or not the defomation state is bending dominated. Under certain special shapes of the load, such as the constant pressure shape, the structure is solid in spite of missing support, and the test fails. In that case, referred to as the soft membrane-dominated case, the right scaling of the load to achieve a nontrivial asymptotic state lets the load be proportional to t, i.e., the same loading as in the hard membrane case. For a more general classification of shell asymptotics, the reader is referred to the introductory part of [P].

We denote the displacement fields corresponding to the above five models as $\underline{U}^{3D}$, $\underline{U}^R$, $\underline{U}^K$, $\underline{U}^0$, and $\underline{U}^M$, respectively. The main results of the paper are the estimates

$$|||\underline{U}^{3D} - \underline{U}^R|||_{3D} = \mathcal{O}(\sigma^{1/2} t^{1/2}), \tag{1.1a}$$

$$|||\underline{U}^R - \underline{U}^K|||_{3D} = \mathcal{O}(\sigma t^{1/2}), \tag{1.1b}$$

$$|||\underline{U}^K - \underline{U}^0|||_{3D} = \mathcal{O}(t^{1/4}), \tag{1.1c}$$

$$|||\underline{U}^K - \underline{U}^M|||_{3D} = \mathcal{O}(t^{5/4}), \tag{1.1d}$$

where $|||\cdot|||_{3D}$ is the relative energy norm corresponding to the three-dimensional model, scaled so that $|||\underline{U}^{3D}|||_{3D}$ is uniformly bounded away from zero and infinity as $t \to 0$, and

$$\sigma = \begin{cases} 1 & \text{in the bending-dominated case,} \\ t & \text{in the soft membrane case.} \end{cases} \tag{1.2}$$

We also show that (1.1c) and (1.1d) are the best possible estimates in the general case.

The plan of the paper is as follows: Starting from the basic formulation of the shell models in §2, we proceed to analyze the behavior of $\underline{U}^0$, $\underline{U}^M$, and $\underline{U}^K$ in more detail in §§3 and 4. The main results (1.1) are then proved in §§5, 6, and 7. In the Appendix we expand $\underline{U}^K - \underline{U}^0$ in more detail in case of two special load distributions of bending-dominated type. Here it turns out that the interior convergence rate is somewhat faster than the global one.

Standard Sobolev space notation (see [PP]) is used throughout the paper. We denote by C or c various constants taking different values on different usage. The constants are independent of parameter t except when indicated explicitly.

2. The shell models. We will retain the notation of [PP] whenever possible. However, to make this paper readable without references, we repeat the main notation to be used in what follows.

We work in the cylindrical coordinate system $(\alpha_1, \alpha_2, \alpha_3)$, where the shell occupies the region

$$\Omega = \left\{ (\alpha_1, \alpha_2, \alpha_3) \in R^3 \mid (\alpha_1, \alpha_2) \in \omega,\ -\frac{t}{2} < \alpha_3 < \frac{t}{2} \right\}.$$

Here ω stands for the midsurface of the shell with

$$\omega = \{(\alpha_1, \alpha_2) \in R^2 \mid -1 < \alpha_1 < 1,\ -\pi < \alpha_2 < \pi\}.$$

According to the three-dimensional elastic model, the displacement field $\underline{U}^{3D} = (U_1^{3D}, U_2^{3D}, U_3^{3D})$ minimizes, in the given energy space $\mathcal{U}^{3D}$, the total energy

$$F^{3D}(\underline{U}) = \frac{1}{2}\mathcal{A}^{3D}(\underline{U}, \underline{U}) - Q^{3D}(\underline{U}), \tag{2.1}$$

where $\mathcal{A}^{3D}(\underline{U}, \underline{V})$ is a bilinear form defined as

$$\mathcal{A}^{3D}(\underline{U}, \underline{V}) = D^{-1} \int_\Omega \left\{ \lambda\, \mathrm{tr}\underline{\underline{e}}(\underline{U})\mathrm{tr}\underline{\underline{e}}(\underline{V}) + \mu \sum_{i,j=1}^{3} e_{ij}(\underline{U}) e_{ij}(\underline{V}) \right\} \chi^{-1}\, d\underline{\alpha},$$

and, furthermore, $\underline{\underline{e}}(\underline{V}) = \{e_{ij}\}_{i,j=1}^3$ is the strain tensor corresponding to a displacement field $\underline{V}$ such that

$$\begin{aligned} e_{11} &= V_{1,1}, & e_{22} &= \chi(V_{2,2} + V_3), \\ e_{12} &= \frac{1}{2}\big(\chi V_{1,2} + V_{2,1}\big), & e_{23} &= \frac{1}{2}\big(V_{2,3} + \chi(V_{3,2} - V_2)\big), \\ e_{13} &= \frac{1}{2}(V_{1,3} + V_{3,1}), & e_{33} &= V_{3,3}. \end{aligned}$$

Here $V_{i,j}$ stands for $\partial V_i/\partial\alpha_j$, $\chi = 1/(1+\alpha_3)$, and $d\underline{\alpha}$ is the abbreviation for $d\alpha_1\, d\alpha_2\, d\alpha_3$. Furthermore, λ and μ are material parameters depending on the Young modulus $E > 0$ and the Poisson ratio ν, $0 \le \nu < 1/2$, and D is a scaling factor. These are defined as follows:

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{E}{1+\nu}, \qquad D = \frac{E\sigma^{-2}t^3}{12(1-\nu^2)},$$

where σ is now and in what follows the same as in (1.2). In (2.1) the quadratic part $\mathcal{A}^{3D}(\underline{U}, \underline{U})$ represents the deformation energy and the linear part

$$Q^{3D}(\underline{V}) = \int_\omega f(\alpha_1, \alpha_2) V_3 \left(\alpha_1, \alpha_2, \frac{t}{2}\right)\left(1 + \frac{t}{2}\right) d\alpha_1\, d\alpha_2$$

is the potential energy due to the external load $F(\alpha_1, \alpha_2) = D \cdot f(\alpha_1, \alpha_2)$. As mentioned in §1, we assume that the actual load F is proportional to t^3 or t depending on the deformation state, so the scaled load f is independent of t.

The dimensionally reduced models are, in general, derived from the assumption that the displacement field is a low-order polynomial of α_3 for each (α_1, α_2). Here we assume that

$$\begin{aligned} U_1(\alpha_1, \alpha_2, \alpha_3) &= u(\alpha_1, \alpha_2) - \alpha_3\theta_1(\alpha_1, \alpha_2), \\ U_2(\alpha_1, \alpha_2, \alpha_3) &= v(\alpha_1, \alpha_2) - \alpha_3\theta_2(\alpha_1, \alpha_2), \\ U_3(\alpha_1, \alpha_2, \alpha_3) &= w(\alpha_1, \alpha_2) + \alpha_3\psi_1(\alpha_1, \alpha_2) + \alpha_3^2\psi_2(\alpha_1, \alpha_2)/2. \end{aligned} \tag{2.2}$$

In the Reissner–Mindlin-type model, $\underline{u}^R = (u^R, v^R, w^R, \theta_1^R, \theta_2^R)$ is obtained by minimizing the energy

$$F^R(\underline{u}) = \frac{\sigma^2}{2}\left\{\mathcal{A}^R(\underline{u}, \underline{u}) + t^{-2}\mathcal{B}^R(\underline{u}, \underline{u}) + t^{-2}\mathcal{C}^R(\underline{u}, \underline{u})\right\} - q(\underline{u}) \tag{2.3}$$

in a certain energy space $\mathcal{U}^R$ (see below), where the bilinear forms $\mathcal{A}^R(\underline{u},\underline{v})$, $\mathcal{B}^R(\underline{u},\underline{v})$, and $\mathcal{C}^R(\underline{u},\underline{v})$ represent the scaled bending, membrane, and transverse shear energy, respectively. These are defined as

$$\mathcal{A}^R(\underline{u},\underline{\tilde{v}}) = \int_\omega \left(\nu \operatorname{tr}\underline{\underline{\kappa}} \operatorname{tr}\underline{\underline{\tilde{\kappa}}} + (1-\nu) \sum_{i,j=1}^{2} \kappa_{ij}\tilde{\kappa}_{ij} \right) d\underline{\alpha}, \tag{2.4a}$$

$$\mathcal{B}^R(\underline{u},\underline{\tilde{v}}) = 12 \int_\omega \left(\nu \operatorname{tr}\underline{\underline{\beta}} \operatorname{tr}\underline{\underline{\tilde{\beta}}} + (1-\nu) \sum_{i,j=1}^{2} \beta_{ij}\tilde{\beta}_{ij} \right) d\underline{\alpha}, \tag{2.4b}$$

$$\mathcal{C}^R(\underline{u},\underline{\tilde{v}}) = 6(1-\nu) \int_\omega (\rho_1\tilde{\rho}_1 + \rho_2\tilde{\rho}_2)\, d\underline{\alpha}, \tag{2.4c}$$

where, furthermore $\underline{\underline{\kappa}} = \{\kappa_{ij}\}_{i,j=1}^2$ and $\underline{\underline{\beta}} = \{\beta_{ij}\}_{i,j=1}^2$ are the dimensionally reduced bending and membrane strains

$$\begin{aligned} \beta_{11} &= u_{,1}\,, & \kappa_{11} &= \theta_{1,1}, \\ \beta_{12} &= \frac{1}{2}(u_{,2} + v_{,1}\,), & \kappa_{12} &= \frac{1}{2}(\theta_{1,2} + \theta_{2,1} - v_{,1}\,), \\ \beta_{22} &= v_{,2} + w, & \kappa_{22} &= \theta_{2,2}, \end{aligned}$$

and ρ_i are the transverse shear strains

$$\rho_1 = -\theta_1 + w_{,1}\,, \qquad \rho_2 = -\theta_2 + w_{,2} - v.$$

The auxiliary functions ψ_1 and ψ_2 in (2.2) are related to the membrane and bending strains as

$$\psi_1 = -\frac{\nu}{1-\nu}\operatorname{tr}\underline{\underline{\beta}}, \qquad \psi_2 = \frac{\nu}{1-\nu}\operatorname{tr}\underline{\underline{\kappa}}. \tag{2.5}$$

These expressions, derived from the energy principle [PP], are assumed in all dimension reduction models considered whenever expansion (2.2) is used in the three-dimensional interpretation of the displacement fields. Finally, the dimensionally reduced potential energy in (2.3) has the form

$$q(\underline{u}) = \int_\omega f \cdot w \, d\underline{\alpha}. \tag{2.6}$$

The Kirchhoff-type model is achieved by eliminating rotations (θ_1, θ_2). This is done by imposing the so called Kirchhoff–Love constraints $\rho_1 = \rho_2 = 0$, i.e.,

$$\theta_1 = w_{,1}\,, \qquad \theta_2 = w_{,2} - v. \tag{2.7}$$

Hence, $\underline{u}^K = (u^K, v^K, w^K)$ minimizes, in the given displacement space $\mathcal{U}^K$, the energy

$$F^K(\underline{u}) = \frac{\sigma^2}{2}\left\{\mathcal{A}^K(\underline{u},\underline{u}) + t^{-2}\mathcal{B}^K(\underline{u},\underline{u})\right\} - q(\underline{u}), \tag{2.8}$$

where $\mathcal{A}^K$, $\mathcal{B}^K$, and q are defined as in (2.4a), (2.4b), and (2.6), respectively. The membrane strains are here defined as above and the bending strains are now written as

$$\kappa_{11} = w_{,11}\,, \qquad \kappa_{12} = w_{,12} - v_{,1}\,, \qquad \kappa_{22} = w_{,22} - v_{,2}\,.$$

The auxiliary functions $(\psi_1, \psi_2) = (\psi_1^K, \psi_2^K)$ are again defined by (2.5).

The asymptotic *inextensional* shell model is obtained by taking a formal limit of $\underline{u}^K$ as $t \to 0$ with $\sigma = 1$. This means that $\underline{u}^0 = (u^0, v^0, w^0)$ minimizes the reduced energy

$$F^0(\underline{u}) = \frac{1}{2}\mathcal{A}^K(\underline{u}, \underline{u}) - q(\underline{u})$$

in $\mathcal{U}^0 = \{\underline{u} \in \mathcal{U}^K \mid \mathcal{B}^K(\underline{u}, \underline{u}) = 0\}$, and $(\theta_1, \theta_2, \psi_1, \psi_2) = (\theta_1^0, \theta_2^0, \psi_1^0, \psi_2^0)$ are defined by (2.7) and (2.5).

Finally, the asymptotic *membrane theory* is obtained by taking a formal limit of $\underline{u}^K$ as $t \to 0$, but this time we assume that $\sigma = t$ in (2.8). The limit solution $\underline{u}^M$ then minimizes the reduced energy

$$F^M(\underline{u}) = \frac{1}{2}\mathcal{B}^K(\underline{u}, \underline{u}) - q(\underline{u})$$

in a certain energy space $\mathcal{U}^M$ defined below.

The (scaled) energy norms in the nonasymptotic models are defined in the usual way, namely,

$$\begin{aligned}
|||\underline{U}|||_{3D} &= \sqrt{\mathcal{A}^{3D}(\underline{U}, \underline{U})}, \\
|||\underline{u}|||_{R,t} &= \sigma\sqrt{\mathcal{A}^R(\underline{u}, \underline{u}) + t^{-2}\mathcal{B}^R(\underline{u}, \underline{u}) + t^{-2}\mathcal{C}^R(\underline{u}, \underline{u})}, \\
|||\underline{u}|||_{K,t} &= \sigma\sqrt{\mathcal{A}^K(\underline{u}, \underline{u}) + t^{-2}\mathcal{B}^K(\underline{u}, \underline{u})},
\end{aligned}$$

and in the asymptotic models as

$$\begin{aligned}
|||\underline{u}|||_0 &= \sqrt{\mathcal{A}^K(\underline{u}, \underline{u})} \quad \text{(inextensional theory)}, \\
|||\underline{u}|||_M &= \sqrt{\mathcal{B}^K(\underline{u}, \underline{u})} \quad \text{(membrane theory)}.
\end{aligned}$$

The aim of the remaining part of this section is to define the energy spaces $\mathcal{U}^{3D}$, $\mathcal{U}^R$, $\mathcal{U}^K$, $\mathcal{U}^0$, $\mathcal{U}^M$ and give some coersivity results (without proofs). These results guarantee the existence and uniqueness of the displacement fields presented above.

We begin with the three-dimensional case. Let

$$\mathcal{W}^{3D} = \{\underline{U} \in H^1(\Omega)^3 \mid \underline{U}(\alpha_1, -\pi, \alpha_3) = \underline{U}(\alpha_1, \pi, \alpha_3)\}$$

and q_i^{3D}, $i = 1, \ldots, 6$ be functionals defined on $\mathcal{W}^{3D}$ such that

$$\begin{aligned}
q_1^{3D}(\underline{U}) &= \int_\Omega U_1 \, d\underline{\alpha}, \\
q_2^{3D}(\underline{U}) &= \int_\Omega (U_2 \cdot \alpha_1 \sin\alpha_2 - U_3 \cdot \alpha_1 \cos\alpha_2) \, d\underline{\alpha}, \\
q_3^{3D}(\underline{U}) &= \int_\Omega (U_2 \cdot \alpha_1 \cos\alpha_2 + U_3 \cdot \alpha_1 \sin\alpha_2) \, d\underline{\alpha}, \\
q_4^{3D}(\underline{U}) &= \int_\Omega (U_2 \cdot \cos\alpha_2 + U_3 \cdot \sin\alpha_2) \, d\underline{\alpha}, \\
q_5^{3D}(\underline{U}) &= \int_\Omega (U_2 \cdot \sin\alpha_2 - U_3 \cdot \cos\alpha_2) \, d\underline{\alpha}, \\
q_6^{3D}(\underline{U}) &= \int_\Omega U_2 \, d\underline{\alpha}.
\end{aligned} \tag{2.9}$$

Then, defining the energy space $\mathcal{U}^{3D}$ as

$$\mathcal{U}^{3D} = \left\{ \underline{U} \in \mathcal{W}^{3D} |\ q_i^{3D}(\underline{U}) = 0,\ i = 1, \dots, 6 \right\},$$

we have the following coersivity result (see [NH]):

$$\mathcal{A}^{3D}(\underline{U}, \underline{U}) \geq c(t)\ ||\underline{U}||^2_{1,\Omega}, \qquad \underline{U} \in \mathcal{U}^{3D}. \tag{2.10}$$

We point out that the space

$$\mathcal{P} = \left\{ \underline{U} \in \mathcal{W}^{3D} |\ \mathcal{A}^{3D}(\underline{U}, \underline{U}) = 0 \right\}$$

consists precisely of the rigid displacements of the shell (cf. [NH]), namely,

$$\begin{aligned} \mathcal{P} = \{ \underline{U} \in \mathcal{W}^{3D} | \\ & U_1 = c_1 + c_2(1+\alpha_3)\cos\alpha_2 - c_3(1+\alpha_3)\sin\alpha_2, \\ & U_2 = -c_4(1+\alpha_3) + (c_5 + c_3\alpha_1)\cos\alpha_2 - (c_6 - c_2\alpha_1)\sin\alpha_2, \\ & U_3 = (c_6 - c_2\alpha_1)\cos\alpha_2 + (c_5 + c_3\alpha_1)\sin\alpha_2, \\ & c_i \in R,\ i = 1, \dots, 6\ \}. \end{aligned} \tag{2.11}$$

Now consider the variational problem of finding $\underline{U} \in \mathcal{U}^{3D}$ such that

$$\mathcal{A}^{3D}(\underline{U}^{3D}, \underline{V}) = Q^{3D}(\underline{V}), \qquad \text{for any } \underline{V} \in \mathcal{U}^{3D}.$$

By (2.10) and by the Riesz representation theorem, this problem is uniquely solvable. Moreover, since $\mathcal{W}^{3D} = \mathcal{U}^{3D} + \mathcal{P}$, $\underline{U}^{3D}$ minimizes F^{3D} in $\mathcal{W}^{3D}$ provided that the load is such that $Q^{3D}(\underline{V}) = 0$ for any $\underline{V} \in \mathcal{P}$, or, equivalently,

$$\begin{aligned} 0 &= \int_\omega f(\alpha_1, \alpha_2) \cdot \cos\alpha_2\ d\underline{\alpha} = \int_\omega f(\alpha_1, \alpha_2) \cdot \sin\alpha_2\ d\underline{\alpha} \\ &= \int_\omega f(\alpha_1, \alpha_2) \cdot \alpha_1 \cos\alpha_2\ d\underline{\alpha} = \int_\omega f(\alpha_1, \alpha_2) \cdot \alpha_1 \sin\alpha_2\ d\underline{\alpha}. \end{aligned} \tag{2.12}$$

We assume below that this equilibrium condition holds. It is also convenient to summarize in this context all the other constraints that f should satisfy. Hence, in the bending-dominated case we assume that

(a) $f \in C^\infty(\overline{\omega})$,
(b) $(\partial^j f/\partial\alpha_2^j)(\alpha_1, -\pi) = (\partial^j f/\partial\alpha_2^j)(\alpha_1, \pi)$, $j = 0, 1, 2, \dots$,
(c) f is independent of t,
(d) $G_1(\alpha_2) := \int_I f_{,2}(\alpha_1, \alpha_2)\ d\alpha_1 \neq 0$ or $G_2(\alpha_2) := \int_I \alpha_1 f_{,22}(\alpha_1, \alpha_2)\ d\alpha_1 \neq 0$,
(e) f satisfies (2.12),

where (here and in what follows) $I = (-1, 1)$. Note that (a) and (b) hold if and only if f is a restriction to ω of a smooth function $\tilde{f}$ defined on R^2, and is 2π-periodic in α_2. The set of such functions is denoted by $C^\infty_{\text{per}}(\overline{\omega})$. Furthermore, it turns out (see the next section) that in the bending-dominated case $\underline{u}^0 = 0$; if and only if $G_1 = G_2 = 0$, i.e., assumption (d) is equivalent to the assumption that the deformation state of the shell is bending dominated. Hence, the soft membrane case is obtained if f satisfies (a), (b), (c), and

$$G_1(\alpha_2) = G_2(\alpha_2) = 0. \tag{d'}$$

Note that (d′) implies (e).

The dimensionally reduced cases are handled almost similarly. Let

$$\begin{aligned}\mathcal{W}^R &= \{(u,v,w,\theta_1,\theta_2) \in H^1(\omega)^5 \mid \\ &\qquad (u,v,w,\theta_1,\theta_2)(\alpha_1,-\pi) = (u,v,w,\theta_1,\theta_2)(\alpha_1,\pi)\},\\ \mathcal{W}^K &= \{(u,v,w) \in H^1(\omega)^2 \times H^2(\omega) \mid \\ &\qquad (u,v,w,w_{,2})(\alpha_1,-\pi) = (u,v,w,w_{,2})(\alpha_1,\pi)\},\end{aligned}$$

and let q_i^R, q_i^K, $i = 1,\dots,6$, be functionals defined in $\mathcal{W}^R$ and $\mathcal{W}^K$, which are otherwise of the form (2.9), but where (U_1,U_2,U_3) is now replaced by (u,v,w) and Ω is replaced by ω. We further define the following energy spaces:

$$\begin{aligned}\mathcal{U}^R &= \{\underline{u} \in \mathcal{W}^R \mid q_i^R(\underline{u}) = 0,\ i = 1,\dots,6\},\\ \mathcal{U}^K &= \{\underline{u} \in \mathcal{W}^K \mid q_i^K(\underline{u}) = 0,\ i = 1,\dots,6\},\\ \mathcal{U}^0 &= \{\underline{u} \in \mathcal{U}^K \mid \mathcal{B}^K(\underline{u},\underline{u}) = 0\}.\end{aligned}$$

Finally, to define the limit energy space $\mathcal{U}^M$ in the soft membrane case, we denote by $(\mathcal{U}^0)^\perp$ the orthogonal complement of $\mathcal{U}^0$ in $\mathcal{U}^K$ with respect to the inner product $(\underline{u},\underline{v}) \mapsto \mathcal{A}^K(\underline{u},\underline{v}) + \mathcal{B}^K(\underline{u},\underline{v})$. Then $\underline{u} \mapsto |||\underline{u}|||_M = \mathcal{B}^K(\underline{u},\underline{u})^{1/2}$ is a norm in $(\mathcal{U}^0)^\perp$. We then define $\mathcal{U}^M$ as the closure of $(\mathcal{U}^0)^\perp$ with respect to this norm.

After these definitions, the dimensionally reduced shell models may be given variational formulations as follows: find $\underline{u}^R \in \mathcal{U}^R$, $\underline{u}^K \in \mathcal{U}^K$, $\underline{u}^0 \in \mathcal{U}^0$, and $\underline{u}^M \in \mathcal{U}^M$ such that

$$\sigma^2\{\mathcal{A}^R(\underline{u}^R,\underline{v}) + t^{-2}\mathcal{B}^R(\underline{u}^R,\underline{v}) + t^{-2}\mathcal{C}^R(\underline{u}^R,\underline{v})\} = q(\underline{v}), \quad \underline{v} \in \mathcal{U}^R, \tag{2.13}$$

$$\sigma^2\{\mathcal{A}^K(\underline{u}^K,\underline{v}) + t^{-2}\mathcal{B}^K(\underline{u}^K,\underline{v})\} = q(\underline{v}), \quad \underline{v} \in \mathcal{U}^K, \tag{2.14}$$

$$\mathcal{A}^K(\underline{u}^0,\underline{v}) = q(\underline{v}), \quad \underline{v} \in \mathcal{U}^0, \tag{2.15}$$

$$\mathcal{B}^K(\underline{u}^M,\underline{v}) = q(\underline{v}), \quad \underline{v} \in \mathcal{U}^M. \tag{2.16}$$

THEOREM 2.1. *Problems* (2.13)–(2.16) *are uniquely solvable. Moreover,*

$$\|\underline{u}^{R,K}\|_{1,\omega} + \|w^K\|_{2,\omega} = \mathcal{O}(\sigma^{-1}), \qquad \|\beta_{ij}^{R,K}\|_{L_2(\omega)} + \|\rho_i^R\|_{L_2(\omega)} = \mathcal{O}(t/\sigma).$$

Proof. Proceeding as in the proof of (2.10) (see [NH]), we find that

$$\mathcal{A}^R(\underline{u},\underline{u}) + \mathcal{B}^R(\underline{u},\underline{u}) + \mathcal{C}^R(\underline{u},\underline{u}) \geq c\|\underline{u}\|_{1,\omega}^2, \quad \underline{u} \in \mathcal{U}^R, \tag{2.17}$$

$$\mathcal{A}^K(\underline{u},\underline{u}) + \mathcal{B}^K(\underline{u},\underline{u}) \geq c\{\|u\|_{1,\omega}^2 + \|v\|_{1,\omega}^2 + \|w\|_{2,\omega}^2\}, \quad \underline{u} \in \mathcal{U}^K, \tag{2.18}$$

so the unique solvability of problems (2.13) and (2.14) follows from the Riesz representation theorem. Note also that since $\mathcal{U}^0$ is a closed subspace of $\mathcal{U}^K$, (2.15) is uniquely solvable whenever (2.14) is solvable.

The unique solvability of (2.16) follows from the Riesz representation theorem as well, once it is shown that q defines a bounded linear functional in $\mathcal{U}^M$. Integrating by parts and applying the constraint (d′) above this follows:

$$\begin{aligned}|q(\underline{u})| &= \left|\int_\omega fw\, d\underline{\alpha}\right| = \left|\int_\omega \{f\beta_{22} - 2I_1(f_{,2})\beta_{12} + I_2(f_{,22})\beta_{11}\}\, d\underline{\alpha}\right| \\ &\leq C\|f\|_{2,\omega}|||\underline{u}|||_M, \quad \underline{u} = (u,v,w) \in \mathcal{U}^K,\end{aligned} \tag{2.19}$$

where and in what follows,

$$I_n(g)(\alpha_1,\alpha_2)=\int_{-1}^{\alpha_1}\int_{-1}^{x_1}\cdots\int_{-1}^{x_{n-1}} g(x_n,\alpha_2)\,dx_n dx_{n-1}\ldots dx_1. \tag{2.20}$$

Since $\mathcal{U}^K$ is dense in $\mathcal{U}^M$, this part of the assertion follows.

The remaining assertions are just basic stability estimates obtained by applying (2.13), (2.14), (2.17), (2.18), and, in the soft membrane case, (2.19). □

Remark 2.1. Exactly as in the three-dimensional case, we have

$$\mathcal{W}^R=\mathcal{U}^R+\mathcal{P}^R,\quad \mathcal{W}^K=\mathcal{U}^K+\mathcal{P}^K,$$

where $\mathcal{P}^R$ and $\mathcal{P}^K$ are defined otherwise as in (2.11), but where $\mathcal{W}^{3D}$ is now replaced by $\mathcal{W}^R$ or $\mathcal{W}^K$, respectively, and α_3 is replaced by 0. By (2.12), both sides in (2.13) vanish if $\underline{v}\in\mathcal{P}^R$ and, accordingly, (2.13) holds for all $\underline{v}\in\mathcal{W}^R$. Furthermore, by the same argument, (2.14) holds for all $\underline{v}\in\mathcal{W}^K$.

3. The asymptotic displacement fields. In this section we solve $\underline{u}^0$ and $\underline{u}^M$ explicitly and show that

$$\underline{u}^0,\underline{u}^M\in\mathcal{U}^K\cap C^\infty_{\mathrm{per}}(\overline{\omega})^3. \tag{3.1}$$

We further show that $\underline{u}^K$ actually converges toward the asymptotic field in the scaled energy norm $|||\cdot|||_{K,t}$.

First, we note that $\mathcal{U}^0$ may be redefined as

$$\begin{aligned}\mathcal{U}^0=\big\{&\underline{u}=(u,v,w)=\big(\phi(\alpha_2)\,,-\alpha_1\phi'(\alpha_2)-\vartheta(\alpha_2)\,,\alpha_1\phi''(\alpha_2)+\vartheta'(\alpha_2)\big),\\ &(\phi,\vartheta)\in H^4(-\pi,\,\pi)\times H^3(-\pi,\,\pi),\\ &\phi^{(j)}(-\pi)=\phi^{(j)}(\pi),\ j=0,1,2,3,\\ &\vartheta^{(j)}(-\pi)=\vartheta^{(j)}(\pi),\ j=0,1,2,\\ &q_i^K(\underline{u})=0,\quad i=1,\ldots,6\big\}.\end{aligned} \tag{3.2}$$

Let us begin with the bending-dominated case. Seeking for a solution to (2.15) in the above form, we obtain by standard calculus of variations that $\underline{u}=\underline{u}(\phi,\vartheta)\in\mathcal{U}^0$ is the desired solution if it satisfies the Euler equations

$$G_1=2\vartheta^{(6)}+4\vartheta^{(4)}+2\vartheta'', \tag{3.3a}$$

$$G_2=\frac{2}{3}\phi^{(8)}-\left(4(1-\nu)-\frac{4}{3}\right)\phi^{(6)}-\left(8(1-\nu)-\frac{2}{3}\right)\phi^{(4)}-4(1-\nu)\phi'', \tag{3.3b}$$

where G_1 and G_2 are definied as in §2. Then, expanding the load in the Fourier series (see [PP])

$$f=\sum_{k=0}^{\infty}\big(f_c^k(\alpha_1)\cos k\alpha_2+f_s^k(\alpha_1)\sin k\alpha_2\big), \tag{3.4}$$

we find that equilibrium condition (2.12) may be rewritten as

$$\int_I f_c^1\,d\alpha_1=\int_I f_s^1\,d\alpha_1=\int\alpha_1 f_c^1\,d\alpha_1=\int_I\alpha_1 f_s^1\,d\alpha_1=0,$$

and, consequently, G_1 and G_2 have expansions

$$G_1 = \sum_{k=2}^{\infty} k \left(\int_I f_s^k \, d\alpha_1 \cos k\alpha_2 - \int_I f_c^k \, d\alpha_1 \sin k\alpha_2 \right),$$

$$G_2 = -\sum_{k=2}^{\infty} k^2 \left(\int_I \alpha_1 f_c^k \, d\alpha_1 \cos k\alpha_2 + \int_I \alpha_1 f_s^k \, d\alpha_1 \sin k\alpha_2 \right).$$

Also expanding ϕ and ϑ into Fourier series, we easily find that (3.3) holds if

$$\begin{aligned} \vartheta &= \sum_{k=2}^{\infty} \left(\vartheta_c^k \cos k\alpha_2 + \vartheta_s^k \sin k\alpha_2 \right), \\ \phi &= \sum_{k=2}^{\infty} \left(\phi_c^k \cos k\alpha_2 + \phi_s^k \sin k\alpha_2 \right), \end{aligned} \tag{3.5a}$$

where

$$\begin{aligned} \vartheta_c^k &= -\left[2k(k^2-1)^2\right]^{-1} \int_I f_s^k \, d\alpha_1, \\ \vartheta_s^k &= \left[2k(k^2-1)^2\right]^{-1} \int_I f_c^k \, d\alpha_1, \\ \phi_a^k &= -\left[\left(\frac{2}{3} k^2 + 4(1-\nu) \right) (k^2-1)^2 \right]^{-1} \int_I \alpha_1 f_a^k \, d\alpha_1, \quad a = c, s. \end{aligned} \tag{3.5b}$$

By (3.5b) there exists a constant c independent of k such that

$$\begin{aligned} |\vartheta_c^k| &\le ck^{-5} \|f_s^k\|_{0,I}, \qquad |\vartheta_s^k| \le ck^{-5} \|f_c^k\|_{0,I}, \\ |\phi_c^k| &\le ck^{-6} \|f_c^k\|_{0,I}, \qquad |\phi_s^k| \le ck^{-6} \|f_s^k\|_{0,I}, \end{aligned}$$

and since $f \in C^{\infty}_{\mathrm{per}}(\overline{\omega})$, we conclude that (ϕ, ϑ) defined by (3.5) is a smooth, 2π-periodic function. Furthermore, it is obvious that $\underline{u}(\phi, \vartheta) \in \boldsymbol{\mathcal{U}}^0$, and thus by uniqueness, $\underline{u}(\phi, \vartheta) = \underline{u}^0 \in C^{\infty}_{\mathrm{per}}(\overline{\omega})^3$. We also note that $\underline{u}^0 \neq \underline{0}$ if and only if either $G_1 \neq 0$ or $G_2 \neq 0$.

Our next step is to solve $\underline{u}^M$ explicitly in the soft membrane case, where $G_1 = G_2 = 0$ and (2.19) holds. To this end, we first note that by (2.19), $\boldsymbol{\mathcal{U}}^0 \subset \mathrm{Ker}(q)$, so

$$\mathcal{B}^K(\underline{u}^M, \underline{v}) = q(\underline{v}) \quad \text{for any} \quad \underline{v} \in \boldsymbol{\mathcal{U}}^K. \tag{3.6a}$$

Then, set $\underline{u} = \tilde{\underline{u}} + \underline{z}$, where

$$\begin{aligned} \tilde{u} &= \left(12(1-\nu^2)\right)^{-1} \left\{ I_3(f_{,22}) - \nu I_1(f) + \phi \right\}, \\ \tilde{v} &= -\left(12(1-\nu^2)\right)^{-1} \left\{ I_4(f_{,222}) + (2+\nu) I_2(f_{,2}) + \alpha_1 \phi' + \vartheta \right\}, \\ \tilde{w} &= \left(12(1-\nu^2)\right)^{-1} \left\{ I_4(f_{,2222}) + 2I_2(f_{,22}) + f + \alpha_1 \phi'' + \vartheta' \right\}, \end{aligned}$$

where $\underline{z} \in \boldsymbol{\mathcal{U}}^0$ is to be defined and (see (3.4))

$$\begin{aligned} \vartheta = &-\frac{1}{2} \int_I \left(2I_4(f_c^1) - (4+\nu) I_2(f_c^1)\right) \, d\alpha_1 \cos\alpha_2 \\ &+ \frac{1}{2} \int_I \left(2I_4(f_s^1) - (4+\nu) I_2(f_s^1)\right) \, d\alpha_1 \sin\alpha_2, \\ \phi = &\frac{\nu}{2} \int_I I_1(f_c^0) \, d\alpha_1 + \frac{3}{4} \int_I \alpha_1 \left(2I_4(f_c^1) - (4+\nu) I_2(f_c^1)\right) \, d\alpha_1 \cos\alpha_2 \\ &+ \frac{3}{4} \int_I \alpha_1 \left(2I_4(f_s^1) - (4+\nu) I_2(f_s^1)\right) \, d\alpha_1 \sin\alpha_2. \end{aligned}$$

It is easy to check that $\underline{u}$ satisfies the Euler equations

$$\begin{aligned} \beta_{11,1} + \nu\beta_{22,1} + (1-\nu)\beta_{12,2} &= 0, \\ \nu\beta_{11,2} + \beta_{22,2} + (1-\nu)\beta_{12,1} &= 0, \\ 12\nu\beta_{11} + 12\beta_{22} &= f, \end{aligned} \tag{3.6b}$$

and the natural boundary conditions

$$\big(\beta_{11} + \nu\beta_{22}\big)(\pm 1, \alpha_2) = \beta_{12}(\pm 1, \alpha_2) = 0. \tag{3.6c}$$

Consequently, $\underline{u}$ satisfies (3.6a). Moreover, by the above definition of (ϕ, ϑ), $\tilde{\underline{u}} \in \boldsymbol{\mathcal{U}}^K \cap C^\infty_{\mathrm{per}}(\overline{\omega})^3$. Our aim is to choose $\underline{z}$ so that

$$\mathcal{A}^K(\underline{u}, \underline{v}) = 0 \quad \text{for all } \ \underline{v} \in \boldsymbol{\mathcal{U}}^0. \tag{3.7}$$

Taking into account that any $\underline{z} \in \boldsymbol{\mathcal{U}}^0$ is of the form (3.2), we conclude, integrating by parts, that (3.7) holds if $\underline{z} = \underline{z}(\phi, \vartheta)$, where ϕ and ϑ satisfy (3.3) with G_i replaced by F_i, $i = 1, 2$, as defined by

$$\begin{aligned} F_1 &= -\int_I \big((\nu\tilde{\kappa}_{11} + \tilde{\kappa}_{22}),_{222} + (\nu\tilde{\kappa}_{11} + \tilde{\kappa}_{22}),_2\big) d\alpha_1, \\ F_2 &= -\int_I \big\{\alpha_1\big((\nu\tilde{\kappa}_{11} + \tilde{\kappa}_{22}),_{2222} + (\nu\tilde{\kappa}_{11} + \tilde{\kappa}_{22}),_{22}\big) \\ &\qquad - 2(1-\nu)\big(\tilde{\kappa}_{12,222} + \tilde{\kappa}_{12,2}\big)\big\} d\alpha_1. \end{aligned}$$

Obviously, Fourier expansions of F_1 and F_2 contain components with $k \geq 2$ only, so proceeding as above we conclude that $\underline{z}(\phi, \vartheta) \in \boldsymbol{\mathcal{U}}^0 \cap C^\infty_{\mathrm{per}}(\overline{\omega})^3$.

Summing up, we have constructed $\underline{u} \in \boldsymbol{\mathcal{U}}^K \cap C^\infty_{\mathrm{per}}(\overline{\omega})^3$ such that (3.6a) holds and, moreover, $\underline{u} \in \big(\boldsymbol{\mathcal{U}}^0\big)^\perp$. By the density of $\boldsymbol{\mathcal{U}}^K \cap \boldsymbol{\mathcal{U}}^M$ in $\boldsymbol{\mathcal{U}}^M$, (3.6a) holds with $\underline{u}^M$ and $\boldsymbol{\mathcal{U}}^K$ replaced by $\underline{u}$ and $\boldsymbol{\mathcal{U}}^M$, and thus, by uniqueness, $\underline{u} = \underline{u}^M$. Accordingly, (3.1) also holds in the soft membrane case.

THEOREM 3.1. *As* $t \to 0$, $|||\underline{u}^K - \underline{u}^0|||_{K,t} \to 0$ (*bending-dominated case*), *or* $t^{-1}|||\underline{u}^K - \underline{u}^M|||_{K,t} \to 0$ (*soft membrane case*).

Proof. We first consider the bending-dominated case.

Let $\varphi(\underline{v}) = q(\underline{v}) - \mathcal{A}^K(\underline{u}^0, \underline{v})$ and $\underline{s}^t = \underline{u}^K - \underline{u}^0$. We note that

$$|||\underline{s}^t|||^2_{K,t} = \varphi(\underline{s}^t), \tag{3.8}$$

and by (2.15),

$$\varphi(\underline{v}) = 0 \quad \text{for any } \ \underline{v} \in \boldsymbol{\mathcal{U}}^0. \tag{3.9}$$

Furthermore, it follows from Theorem 2.1 that

$$\mathcal{A}^K(\underline{s}^t, \underline{s}^t) + \mathcal{B}^K(\underline{s}^t, \underline{s}^t) \ \leq \ C \quad \text{for all } \ t \in (0, t_0). \tag{3.10}$$

By (3.10), any subset of $\{\underline{s}^t,\ 0 < t \leq t_0\}$ contains a subsequence $\{\underline{s}^{t_r}\}$ converging weakly to $\underline{s} \in \boldsymbol{\mathcal{U}}^K$ in the sense that

$$\mathcal{B}^K(\underline{s}^{t_r}, \underline{v}) \to \mathcal{B}^K(\underline{s}, \underline{v}), \quad \mathcal{A}^K(\underline{s}^{t_r}, \underline{v}) \to \mathcal{A}^K(\underline{s}, \underline{v}), \quad \text{and} \tag{3.11a}$$

$$\varphi(\underline{s}^{t_r}) \to \varphi(\underline{s}) \quad \text{as} \quad t_r \to 0 \tag{3.11b}$$

for any $\underline{v} \in \mathcal{U}^K$. Then, by Theorem 2.1,

$$\left|\mathcal{B}^K(\underline{s}^{t_r}, \underline{s})\right| \leq \mathcal{B}^K(\underline{s}^{t_r}, \underline{s}^{t_r})^{1/2}\mathcal{B}^K(\underline{s}, \underline{s})^{1/2} = \mathcal{O}(t_r), \tag{3.12}$$

so $\underline{s} \in \mathcal{U}^0$ by (3.11a) and, accordingly, $\varphi(\underline{s}) = 0$ by (3.9). Then, by (3.8) and (3.11b),

$$|||\underline{s}^{t_r}|||_{K,t_r} \to 0 \quad \text{as } t_r \to 0. \tag{3.13}$$

It follows that $|||s^t|||_{K,t} \to 0$ as $t \to 0$; otherwise we could find a subsequence satisfying (3.11) but violating (3.13).

Next, let us prove the assertion in the soft membrane case. Using the same notation as above, it follows from (2.14), (2.16), and (3.1) that

$$|||\underline{s}^t|||_{K,t}^2 = -t^2\mathcal{A}^K(\underline{u}^M, \underline{s}^t) \leq Ct|||\underline{s}^t|||_{K,t}, \tag{3.14}$$

so that (3.10) also holds this time. Accordingly, we can again extract a weakly convergent subsequence $\{\underline{s}^{t_r}\}$ in the sense that (3.11a) holds for some $\underline{s} \in \mathcal{U}^K$. Furthermore, by (3.14), (3.12) also holds and thus, by (3.11a), $\underline{s} \in \mathcal{U}^0$. Then, since $\mathcal{A}^K(\underline{s}^t, \underline{v}) = 0$ for any $\underline{v} \in \mathcal{U}^0$, it follows from (3.11a) that $\mathcal{A}^K(\underline{s}, \underline{s}) = 0$. Hence, $\underline{s} = 0$, and by (3.14),

$$t_r^{-2}|||\underline{s}^{t_r}|||_{K,t_r}^2 = -\mathcal{A}^K(\underline{u}^M, \underline{s}^{t_r}) \to -\mathcal{A}^K(\underline{u}^M, \underline{s}) = 0 \quad \text{as } t_r \to 0.$$

The last estimate, however, must hold also with t_r replaced by t by the same arguments as above. □

Remark 3.1. Assume the soft membrane case. If we only know that $f \in H^2(\omega)$, then $\underline{u}^M \notin \mathcal{U}^K$ in general. In this case one can still show (see [PP, Thm. 3.1]) that

$$|||\underline{u}^K - \underline{u}^M|||_M^2 + t^2\mathcal{A}^K(\underline{u}^K, \underline{u}^K) \to 0 \quad \text{as } t \to 0,$$

so, in particular, the membrane energy still dominates asymptotically.

4. Regularity of $\underline{u}^K$: The bending-dominated case. In this section we study the nature of the solution $\underline{u}^K$ when the deformation state is bending dominated. We begin with some regularity results, which are consequences of the a priori estimates presented in Theorem 2.1. We point out that (4.1a)–(4.1c) are in fact not sharp; they will be improved in §5.1.

Here and in the subsequent sections we use the abbreviation $\mathcal{O}(\phi(t))$ for a quantity bounded in absolute value by $c(p) \cdot \phi(t) \cdot \|f\|_{p,\omega}$ or by $c(p,k) \cdot \phi(t) \cdot \|f^k\|_{p,I}$ in cases where one Fourier component of f is considered. Here p is some finite integer and the dependence on parameter k is assumed to be algebraic, i.e., there exists $m = m(p) \in N$ such that $c(p,k) \leq c(p)k^m$.

THEOREM 4.1. *$\underline{u}^K \in C^\infty_{\mathrm{per}}(\overline{\omega})^3$, and for any multi-index $\tau = (\tau_1, \tau_2)$,*

$$\|D^\tau u^K\| = \mathcal{O}(1), \quad \tau_1 = 0, \quad \|D^\tau u^K\| = \mathcal{O}(\eta t^{(3-\tau_1)/2}), \quad \tau_1 \geq 1, \tag{4.1a}$$

$$\|D^\tau v^K\| = \mathcal{O}(1), \quad \tau_1 \leq 1, \quad \|D^\tau v^K\| = \mathcal{O}(\eta t^{(4-\tau_1)/2}), \quad \tau_1 \geq 2, \tag{4.1b}$$

$$\|D^\tau w^K\| = \mathcal{O}(1), \quad \tau_1 \leq 1, \quad \|D^\tau w^K\| = \mathcal{O}(\eta t^{(2-\tau_1)/2}), \quad \tau_1 \geq 2, \tag{4.1c}$$

where $\eta = 1$ if $\|\cdot\| = \|\cdot\|_{L_2(\omega)}$ and $\eta = t^{-1/4}$ if $\|\cdot\| = \|\cdot\|_{L_\infty(\omega)}$. Furthermore,

$$D^\tau w^K(\pm 1, \alpha_2) = \mathcal{O}(1) \quad \text{if} \quad \tau_1 \leq 3. \tag{4.2}$$

Remark 4.1. Without loss of generality we may consider below only one Fourier component, so assume that $f(\alpha_1,\alpha_2) = f^k(\alpha_1)\cos k\alpha_2$, $k \geq 0$ (see [PP] and (3.4)). In that case the solution is of the form

$$\begin{aligned} &\big(u(\alpha_1)\cos k\alpha_2,\ v(\alpha_1)\sin k\alpha_2,\ w(\alpha_1)\cos k\alpha_2\big) \quad \text{if } k \geq 1,\\ &\big(u(\alpha_1), w(\alpha_1)\big) \quad \text{if } k = 0, \end{aligned} \tag{4.3}$$

where $\underline{u} = \underline{u}(\alpha_1) \in \mathcal{U}_k^F$ is such that

$$\mathcal{A}_k^F(\underline{u},\underline{\tilde{v}}) + t^{-2}\mathcal{B}_k^F(\underline{u},\underline{\tilde{v}}) = \int_I f^k \cdot \tilde{w}\, d\alpha_1, \quad \underline{\tilde{v}} \in \mathcal{U}_k^F, \tag{4.4}$$

and $\mathcal{A}_k^F(\underline{u},\underline{\tilde{v}})$ and $\mathcal{B}_k^F(\underline{u},\underline{\tilde{v}})$ are otherwise of the form (2.4) and (2.5). Now, however, ω and $d\underline{\alpha}$ are replaced by I and $d\alpha_1$, and

$$\begin{aligned} \beta_{11} &= u', & \kappa_{11} &= w'',\\ \beta_{12} &= \frac{1}{2}(-ku + v'), & \kappa_{12} &= -kw' - v',\\ \beta_{22} &= kv + w, & \kappa_{22} &= -k^2 w - kv \end{aligned}$$

if $k \geq 1$, and

$$\beta_{11} = u',\ \ \beta_{22} = w,\ \ \kappa_{11} = w'',\ \ v = \beta_{12} = \kappa_{12} = \kappa_{22} = 0$$

if $k = 0$. The energy space $\mathcal{U}_k^F$ is naturally of the form

$$\begin{aligned} \mathcal{U}_k^F = \big\{&(u,v,w) \in H^1(I)^2 \times H^2(I) \mid\\ &q_i^K\big(u(\alpha_1)\cos k\alpha_2, v(\alpha_1)\sin k\alpha_2, w(\alpha_1)\cos k\alpha_2\big) = 0,\ i = 1,\dots,6\big\} \end{aligned}$$

if $k \geq 1$ and

$$\mathcal{U}_0^F = \big\{(u,w) \in H^1(I) \times H^2(I) \mid q_i^K(u,0,w) = 0,\ i = 1,\dots,6\big\}.$$

By the same arguments as in Remark 2.1, (4.4) actually holds for all $\underline{v} \in H^1(I)^2 \times H^2(I)$ in the case $k \geq 1$ and for all $\underline{v} \in H^1(I) \times H^2(I)$ in the case $k = 0$. Finally, for later use we denote by $|||\cdot|||_k$ the energy norm that corresponds to (4.4).

Proof of Theorem 4.1. It obviously suffices to prove that the solution $\underline{u}$ to (4.4) satisfies (4.1) and (4.2) with D^τ replaced by $\partial^{\tau_1}/\partial\alpha_1^{\tau_1}$.

We begin with the case where $\|\cdot\| = \|\cdot\|_{L_2(\omega)}$. Below we simply replace v by 0 in the case $k = 0$. By Theorem 2.1,

$$\|u\|_{1,I} + \|v\|_{1,I} + \|w\|_{2,I} = \mathcal{O}(1) \quad \text{and} \quad \|\beta_{ij}\|_{L_2(I)} = \mathcal{O}(t),$$

and by (4.4), $\underline{u}$ satisfies the Euler equations

$$u'' = \frac{1}{2}\big\{k^2(1-\nu)u - k(1+\nu)v' - 2\nu w'\big\}, \tag{4.5a}$$

$$v'' = c\big\{6\big((1+\nu)u' + 2(kv+w)\big) - t^2\big((2-\nu)w'' - k^2 w - kv\big)\big\}, \tag{4.5b}$$

$$w^{(4)} = -12t^{-2}\big(\nu u' + kv + w\big) + 2k^2 w'' - k^4 w + k(2-\nu)v'' - k^3 v + f^k \tag{4.5c}$$

with the natural boundary conditions

$$u' + \nu(kv + w) = 0, \tag{4.6a}$$

$$3ku - (3 + t^2)v' - kt^2w' = 0, \tag{4.6b}$$

$$w'' - \nu(k^2w + kv) = 0, \tag{4.6c}$$

$$w^{(3)} - (2 - \nu)(k^2w' + kv') = 0 \tag{4.6d}$$

at both ends, where $c = k/\big((6 + 2t^2)(1 - \nu)\big)$. Noting that (4.5b) can be rewritten in the form

$$v'' = c\big\{6(1 + \nu)\beta_{11} + 12\beta_{22} - t^2\big((2 - \nu)w'' - k^2w - kv\big)\big\},$$

we have $|v|_{2,I} = \mathcal{O}(t)$. Also, (4.5c) can be rewritten as

$$w^{(4)} = -12\nu t^{-2}\beta_{11} - 12t^{-2}\beta_{22} + 2k^2w'' - k^4w + k(2 - \nu)v'' - k^3v + f^k,$$

and thus $|w|_{4,I} = \mathcal{O}(t^{-1})$. We also note that $|w|_{3,I} = \mathcal{O}(t^{-1})$, for we have from boundary condition (4.6d) that

$$w^{(3)}(\alpha_1) = (2 - \nu)\big(k^2w' + kv'\big)(-1) + \int_{-1}^{\alpha_1} w^{(4)}(s)\, ds,$$

and we already know that $v'(-1) = \mathcal{O}(1)$, $w'(-1) = \mathcal{O}(1)$. Hence, we obtain (4.1c) with $\tau_1 \le 4$ by interpolation. Finally, differentiating in (4.5), the proof is completed by induction and interpolation.

Next, let $\|\cdot\| = \|\cdot\|_{L_\infty(I)}$. The first estimates in (4.1) hold because of the Sobolev imbedding theorem. To prove the second part, we assume that $\alpha_1 \ge 0$ (the case where $\alpha_1 < 0$ is handled similarly). For $g = u, v$, or w we have

(4.7)

$$\begin{aligned} |g^{(s)}(\alpha_1)| &= \left| \int_{\alpha_1 - \epsilon}^{\alpha_1} \partial \left(\frac{x - \alpha_1 + \epsilon}{\epsilon} g^{(s)}(x) \right) \right| \\ &\le \left(\int_{\alpha_1 - \epsilon}^{\alpha_1} \epsilon^{-2}\, dx \right)^{1/2} |g|_{s,I} + \left(\int_{\alpha_1 - \epsilon}^{\alpha_1} \epsilon^{-2}(x - \alpha_1 + \epsilon)^2\, dx \right)^{1/2} |g|_{s+1,I} \\ &\le c\left(\epsilon^{-1/2}|g|_{s,I} + \epsilon^{1/2}|g|_{s+1,I}\right) \end{aligned}$$

for any $0 < \epsilon < 1$. The remaining estimates in (4.1) now follow from this inequality by choosing $\epsilon = t^{1/2}$ and recalling the Sobolev–norm estimates already obtained. Then, (4.2) follows from (4.6c) and (4.6d). □

If w is solved from (4.5) with $k \ge 1$, we get a differential equation of the form

$$w^{(8)} - C_6w^{(6)} + C_4w^{(4)} - C_2w^{(2)} + C_0w = \tilde{f}, \tag{4.8}$$

where C_i are nonnegative constants such that

$$C_0 = k^4(k^2 - 1)^2\big(1 + \mathcal{O}(t^2)\big), \qquad C_2 = 4k^2(k^2 - 1)^2\big(1 + \mathcal{O}(t^2)\big),$$
$$C_4 = 12(1 - \nu^2)t^{-2} + \mathcal{O}(k^4), \qquad C_6 = 4k^2\big(1 + \mathcal{O}(t^2)\big),$$

and, furthermore, $\tilde{f} = \big(1 + \mathcal{O}(t^2)\big)(f^k)^{(4)} - \big(2k^2 + \mathcal{O}(t^2)\big)(f^k)'' + \big(k^4 + \mathcal{O}(t^2)\big)f^k$.

The characteristic polynomial for (4.8) consists of even terms only and, therefore, the roots are of the form

$$\eta_1,\quad \overline{\eta}_1,\ -\eta_1,\ -\overline{\eta}_1,\quad \eta_2,\quad \overline{\eta}_2,\ -\eta_2,\ -\overline{\eta}_2,$$

where $(\eta,\overline{\eta})$ is a conjugate pair. Writing $\eta_j = A_j + iB_j$, we find that in the case $k \geq 2$,

$$\begin{aligned}
A_1 &= C_4^{1/4}\cos\left(\frac{1}{2}\arctan\frac{2\sqrt{C_4}}{C_6}\right) + \mathcal{O}(t^{1/2}) = \mathcal{O}(t^{-1/2}),\\
B_1 &= C_4^{1/4}\sin\left(\frac{1}{2}\arctan\frac{2\sqrt{C_4}}{C_6}\right) + \mathcal{O}(t^{1/2}) = \mathcal{O}(t^{-1/2}),\\
A_2 &= \left(\frac{C_0}{C_4}\right)^{1/4}\cos\left(\frac{1}{2}\arctan\frac{2\sqrt{C_0C_4}}{C_2}\right) + \mathcal{O}(t^{3/2}) = \mathcal{O}(t^{1/2}),\\
B_2 &= \left(\frac{C_0}{C_4}\right)^{1/4}\sin\left(\frac{1}{2}\arctan\frac{2\sqrt{C_0C_4}}{C_2}\right) + \mathcal{O}(t^{3/2}) = \mathcal{O}(t^{1/2}).
\end{aligned} \tag{4.9}$$

If $k = 1$, we have $A_2 = B_2 = 0$, since $C_0 = C_2 = 0$ in (4.8) for $k = 1$. Otherwise (4.9) remains valid. It is easy to check that

$$A_1 - B_1 = \mathcal{O}(t^{1/2}), \qquad A_2 - B_2 = (k-1)\cdot\mathcal{O}(t^{3/2}). \tag{4.10}$$

If $k = 0$, (4.5) implies that

$$w^{(5)} + 12t^{-2}(1-\nu^2)w' = f'. \tag{4.11}$$

Here, the roots of the corresponding characteristic polynomial are 0, η, $\overline{\eta}$, $-\eta$, and $-\overline{\eta}$, where, furthermore, $\eta = A(1+i)$, $A = \left(3(1-\nu^2)\right)^{1/4}t^{-1/2}$.

LEMMA 4.1. *There exists a solution w_0 to* (4.8) *or* (4.11) *such that*

$$\|w_0^{(s)}\|_{L_\infty(I)} = \mathcal{O}(1), \quad s \leq 5.$$

Proof. The fundamental sets of (4.8) ($k \geq 1$) or (4.11) ($k = 0$) are

$$\begin{cases}
\left\{e^{\pm\eta_i\alpha_1}, e^{\pm\overline{\eta}_i\alpha_1},\ i = 1,2\right\} & \text{if } k \geq 2,\\
\left\{1, \alpha_1, \alpha_1^2, \alpha_1^3, e^{\pm\eta_1\alpha_1}, e^{\pm\overline{\eta}_1\alpha_1}\right\} & \text{if } k = 1,\\
\left\{1, e^{\pm\eta\alpha_1}, e^{\pm\overline{\eta}\alpha_1}\right\} & \text{if } k = 0.
\end{cases}$$

By variation of constants, we get the particular solutions

$$\begin{aligned}
w_0 &= \operatorname{Re}\left\{\mathcal{O}(t^{7/2})\left(e^{\eta_1\alpha_1}\int_1^{\alpha_1} e^{-\eta_1 x}\tilde{f}(x)\,dx - e^{-\eta_1\alpha_1}\int_{-1}^{\alpha_1} e^{\eta_1 x}\tilde{f}(x)\,dx\right)\right\}\\
&\quad + \operatorname{Re}\left\{\mathcal{O}(t^{1/2})\left(e^{\eta_2\alpha_1}\int_1^{\alpha_1} e^{-\eta_2 x}\tilde{f}(x)\,dx - e^{-\eta_2\alpha_1}\int_{-1}^{\alpha_1} e^{\eta_2 x}\tilde{f}(x)\,dx\right)\right\}\\
&\quad (k \geq 2),\\
w_0 &= \operatorname{Re}\left\{\mathcal{O}(t^{7/2})\left(e^{\eta_1\alpha_1}\int_1^{\alpha_1} e^{-\eta_1 x}\tilde{f}(x)\,dx - e^{-\eta_1\alpha_1}\int_{-1}^{\alpha_1} e^{\eta_1 x}\tilde{f}(x)\,dx\right)\right\}\\
&\quad + \mathcal{O}(t^2)\sum_{i=0}^{3}\alpha_1^i\int_0^{\alpha_1}\sum_{j=0}^{3} C_{ij}x^j\tilde{f}(x)\,dx \quad (k = 1),\\
w_0 &= \operatorname{Re}\left\{\mathcal{O}(t^2)\left(e^{\eta\alpha_1}\int_1^{\alpha_1} e^{-\eta x}(f^0)'(x)\,dx + e^{-\eta\alpha_1}\int_{-1}^{\alpha_1} e^{\eta x}(f^0)'(x)\,dx\right)\right\}\\
&\quad + \mathcal{O}(t^2)\big(f^0(\alpha_1) - f^0(0)\big) \quad (k = 0),
\end{aligned}$$

where C_{ij} are constants of order $\mathcal{O}(1)$ at most. The assertion now follows easily by differentiation. □

In what follows we choose the fundamental sets of (4.8) and (4.11) in a more practical way. If $k \geq 2$, let us set

(4.12)
$$\begin{aligned}
\Psi_0 &= \cos B_2\alpha_1 \cdot \cosh A_2\alpha_1,\\
\Psi_1 &= (2B_2)^{-1} \sin B_2\alpha_1 \cdot \cosh A_2\alpha_1 + (2A_2)^{-1} \cos B_2\alpha_1 \cdot \sinh A_2\alpha_1,\\
\Psi_2 &= (A_2B_2)^{-1} \sin B_2\alpha_1 \cdot \sinh A_2\alpha_1,\\
\Psi_3 &= 3(A_2^3B_2 + A_2B_2^3)^{-1}\big(A_2 \sin B_2\alpha_1 \cdot \cosh A_2\alpha_1 - B_2 \cos B_2\alpha_1 \cdot \sinh A_2\alpha_1\big),\\
\Psi_4 &= 2e^{-A_1} \cos B_1\alpha_1 \cdot \cosh A_1\alpha_1,\\
\Psi_5 &= 2e^{-A_1} \sin B_1\alpha_1 \cdot \cosh A_1\alpha_1,\\
\Psi_6 &= 2e^{-A_1} \sin B_1\alpha_1 \cdot \sinh A_1\alpha_1,\\
\Psi_7 &= 2e^{-A_1} \cos B_1\alpha_1 \cdot \sinh A_1\alpha_1.
\end{aligned}$$

Expanding the first four basis functions into a Taylor series, we have that

$$\|\Psi_i - \alpha_1^i\|_{s,I} = C(s)\mathcal{O}(t^2), \quad s \geq 0,\ i = 0,1,2,3. \tag{4.13}$$

In the case $k = 1$ the fundamental set is otherwise similar, but now

$$\Psi_i = \alpha_1^i, \qquad i = 0,1,2,3. \tag{4.14}$$

Finally, if $k = 0$, we set

$$\begin{aligned}
\Psi_0 &= 1,\\
\Psi_4 &= 2e^{-A} \cos A\alpha_1 \cosh A\alpha_1,\\
\Psi_5 &= 2e^{-A} \sin A\alpha_1 \cosh A\alpha_1,\\
\Psi_6 &= 2e^{-A} \sin A\alpha_1 \sinh A\alpha_1,\\
\Psi_7 &= 2e^{-A} \cos A\alpha_1 \sinh A\alpha_1.
\end{aligned} \tag{4.15}$$

Let us assume for a while that $k \geq 1$. Obviously the solution w we are looking for can be written in the form

$$w = \sum_{i=0}^{7} W^i \cdot \Psi_i + w_0, \qquad W^i \in R, \tag{4.16}$$

where w_0 is as in Lemma 4.1. By (4.2) and Lemma 4.1

$$X_i^{\pm} := w^{(i)}(\pm 1) - w_0^{(i)}(\pm 1) = \mathcal{O}(1), \qquad i = 0,1,2,3. \tag{4.17}$$

The coefficients W^i can then be solved from the linear system

$$\sum_{i=0}^{7} W^i \cdot \Psi_i^{(j)}(\pm 1) = X_j^{\pm}, \qquad j = 0,1,2,3.$$

Setting $Y_i^p := X_i^+ + X_i^-$, $Y_i^m = X_i^+ - X_i^-$, by straightforward computations we have

$$
\begin{aligned}
W^0 &= \ 1/2\ Y_0^p - 1/4\ Y_1^m + \mathcal{O}(t^{1/2}),\\
W^1 &= \ 3/4\ Y_0^m - 1/4\ Y_1^p + \mathcal{O}(t^{1/2}),\\
W^2 &= \ 1/4\ Y_1^m + \mathcal{O}(t^{1/2}),\\
W^3 &= -1/4\ (Y_0^m - Y_1^p) + \mathcal{O}(t^{1/2}),\\
W^4 &= -(2A_1)^{-2}(\cos B_1 - \sin B_1)\big(Y_1^m - Y_2^p\big) + \mathcal{O}(t^{3/2}),\\
W^5 &= \ (2A_1)^{-2}(\cos B_1 + \sin B_1)\big(3Y_0^m - 3Y_1^p + Y_2^m\big) + \mathcal{O}(t^{3/2}),\\
W^6 &= -(2A_1)^{-2}(\cos B_1 + \sin B_1)\big(Y_1^m - Y_2^p\big) + \mathcal{O}(t^{3/2}),\\
W^7 &= \ (2A_1)^{-2}(\cos B_1 - \sin B_1)\big(3Y_0^m - 3Y_1^p + Y_2^m\big) + \mathcal{O}(t^{3/2}),
\end{aligned}
\tag{4.18}
$$

where the residual estimates follow from Theorem 4.1, (4.9), (4.10), and (4.17).

If $k = 0$, we have $\underline{u}^K = (u, w)$, where w is of the form

$$
w = W^0 \cdot \Psi_0 + \sum_{i=4}^{7} W^i \cdot \Psi_i + w_0. \tag{4.19}
$$

By arguments similar to those above, we conclude that

$$
W^0 = \mathcal{O}(1), \qquad W^i = \mathcal{O}(t), \qquad i = 4,5,6,7. \tag{4.20}
$$

THEOREM 4.2. *Let* $\tau = (\tau_1, \tau_2)$. *Then*

$$
D^\tau w^K(\pm 1, \alpha_2) = \mathcal{O}(t^{-1}), \qquad \tau_1 = 4, \tag{4.21}
$$

$$
D^\tau w^K(\pm 1, \alpha_2) = \mathcal{O}(t^{-3/2}), \qquad \tau_1 = 5, \tag{4.22}
$$

$$
D^\tau \beta_{ii}^K(\pm 1, \alpha_2) = \mathcal{O}(t), \qquad \tau_1 = 0,\ i = 1,2, \tag{4.23}
$$

$$
D^\tau \beta_{ii}^K(\pm 1, \alpha_2) = \mathcal{O}(t^{1/2}), \qquad \tau_1 = 1,\ i = 1,2. \tag{4.24}
$$

Proof. We may assume again that $f(\alpha_1, \alpha_2) = f^k(\alpha_1)\cos k\alpha_2$, which implies that $\underline{u}^K$ is of the form (4.3), and apply the analysis above. First, (4.21) and (4.22) follow immediately from (4.16), (4.18), (4.9), (4.19), (4.20), and Lemma 4.1. To prove (4.26) and (4.27), note that (4.5a) and (4.5c) can be rewritten as

$$
\beta_{11}' + \nu\beta_{22}' = -(1-\nu)k\beta_{12}, \tag{4.25}
$$

$$
\nu\beta_{11} + \beta_{22} = -\frac{t^2}{12}\big(w^{(4)} - 2k^2 w^{(2)} + k^4 w - k(2-\nu)v^{(2)} + k^3 v - f^k\big), \tag{4.26}
$$

so that by (4.6a) and (4.26)

$$
\begin{aligned}
&\beta_{11}(\pm 1) + \nu\beta_{22}(\pm 1) = 0,\\
&\nu\beta_{11}(\pm 1) + \beta_{22}(\pm 1) = -\frac{t^2}{12} w^{(4)}(\pm 1) + \mathcal{O}(t^2),
\end{aligned}
$$

where the residual estimate is based on Theorem 4.1. Then (4.23) follows from (4.21) and, furthermore, since by (4.25), (4.6b), (4.26), and Theorem 4.1

$$
\begin{aligned}
&\beta_{11}'(\pm 1) + \nu\beta_{22}'(\pm 1) = -\frac{1+\nu}{6} k t^2 \kappa_{12}(\pm 1) = \mathcal{O}(t^2),\\
&\nu\beta_{11}'(\pm 1) + \beta_{22}'(\pm 1) = -\frac{t^2}{12} w^{(5)}(\pm 1) + \mathcal{O}(t^2),
\end{aligned}
$$

the estimate (4.24) follows as well. □

5. Proofs of (1.1c) and (1.1d).

5.1. Proof of (1.1c): The bending-dominated case. Having established the above preliminaries, we are able to sharpen the result of Theorem 3.1 in the bending-dominated case. We begin with the following theorem.

THEOREM 5.1. $|||\underline{u}^K - \underline{u}^0|||_{K,t} = \mathcal{O}(t^{1/4})$.

Proof. We are going to use the following notation in what follows:

$$(g_1, g_2) = \int_\omega g_1 g_2 \, d\underline{\alpha}, \qquad ((g_1, g_2)) = \int_{-\pi}^{\pi} g_1(1, \alpha_2) g_2(1, \alpha_2) \, d\alpha_2.$$

Furthermore, for $\underline{v} = (u, v, w) \in \mathcal{W}^K$, we denote by r any finite sum of inner products of the form $(\beta_{ij}(\underline{v}), g)$, where g is a smooth function independent of t.

Thus, let $\underline{v} = \underline{u}^K - \underline{u}^0 = (u, v, w) \in \mathcal{U}^K$. Integrating by parts, we conclude first that

$$q(\underline{v}) = ((G_2, u)) - ((G_1', u)) + ((G_1, v)) + r,$$

where q, G_1, and G_2 are defined as in §2. Furthermore, recalling from (3.2) that

$$\kappa_{11}^0 = 0, \quad \kappa_{12}^0 = \phi^{(3)} + \phi', \quad \kappa_{22}^0 = \alpha_1\big(\phi^{(4)} + \phi''\big) + \vartheta^{(3)} + \vartheta',$$

we have

$$\begin{aligned}
&\mathcal{A}^K(\underline{u}^0, \underline{v}) + t^{-2}\mathcal{B}^K(\underline{u}^0, \underline{v}) = \mathcal{A}^K(\underline{u}^0, \underline{v}) \\
&\quad = \big(\alpha_1(\phi^{(4)} + \phi'') + \vartheta^{(3)} + \vartheta', w_{,22}\big) - \big(\alpha_1(\phi^{(4)} + \phi'') + \vartheta^{(3)} + \vartheta', v_{,2}\big) \\
&\qquad + 2(1-\nu)\big(\phi^{(3)} + \phi', w_{,12}\big) - 2(1-\nu)\big(\phi^{(3)} + \phi', v_{,1}\big) \\
&\qquad + \nu\big(\alpha_1(\phi^{(4)} + \phi'') + \vartheta^{(3)} + \vartheta', w_{,11}\big) \\
&\quad = I + II + III + IV + V.
\end{aligned}$$

All the following results are obtained by integrating by parts. First,

$$\begin{aligned}
I &= \big(\alpha_1(\phi^{(6)} + \phi^{(4)}) + \vartheta^{(5)} + \vartheta^{(3)}, w\big) \\
&= -\big(\alpha_1(\phi^{(6)} + \phi^{(4)}) + \vartheta^{(5)} + \vartheta^{(3)}, v_{,2}\big) + r \\
&= \big(\alpha_1(\phi^{(7)} + \phi^{(5)}) + \vartheta^{(6)} + \vartheta^{(4)}, v\big) + r.
\end{aligned}$$

Thus by (3.3b),

$$\begin{aligned}
I + II &= \big(\alpha_1(\phi^{(7)} + 2\phi^{(5)} + \phi^{(3)}) + \vartheta^{(6)} + 2\vartheta^{(4)} + \vartheta'', v\big) + r \\
&= \int_\omega \big(\phi^{(7)} + 2\phi^{(5)} + \phi^{(3)}\big) v \, d\alpha_2 \, \partial_1 \left(\frac{\alpha_1^2}{2} - \frac{1}{2}\right) \\
&\quad + \int_\omega \big(\vartheta^{(6)} + 2\vartheta^{(4)} + \vartheta''\big) v \, d\alpha_2 \, \partial_1(\alpha_1 + 1) + r \\
&= -\left(\phi^{(8)} + 2\phi^{(6)} + \phi^{(4)}, u\left(\frac{\alpha_1^2}{2} - \frac{1}{2}\right)\right) \\
&\quad - \big(\vartheta^{(7)} + 2\vartheta^{(5)} + \vartheta^{(3)}, u(\alpha_1 + 1)\big) + ((G_1, v)) + r \\
&= -\int_\omega \big(\phi^{(8)} + 2\phi^{(6)} + \phi^{(4)}\big) u \, d\alpha_2 \, \partial_1 \left(\frac{\alpha_1^3}{6} - \frac{\alpha_1}{2} - \frac{1}{3}\right) \\
&\quad - \int_\omega \big(\vartheta^{(7)} + 2\vartheta^{(5)} + \vartheta^{(3)}\big) u \, d\alpha_2 \, \partial_1 \left(\frac{\alpha_1^2}{2} + \alpha_1 + \frac{1}{2}\right) + ((G_1, v)) + r \\
&= \frac{2}{3}((\phi^{(8)} + 2\phi^{(6)} + \phi^{(4)}, u)) - ((G_1', u)) + ((G_1, v)) + r.
\end{aligned}$$

Furthermore,

$$\begin{aligned} III &= -2(1-\nu)\big(\phi^{(4)}+\phi'', w_{,1}\big) = -2(1-\nu)\int_{\partial\omega}\big(\phi^{(4)}+\phi''\big)w\ n_1\ ds \\ &= -2(1-\nu)\int_{\partial\omega}\big(\phi^{(5)}+\phi^{(3)}\big)v\ n_1\ ds - 2(1-\nu)\int_{\partial\omega}\big(\phi^{(4)}+\phi^{(2)}\big)\beta_{22}\ n_1\ ds \\ &= -2(1-\nu)\int_{\partial\omega}\big(\phi^{(5)}+\phi^{(3)}\big)v\ n_1\ ds - R, \end{aligned}$$

and thus,

$$\begin{aligned} III + IV &= -2(1-\nu)\int_{\partial\omega}\big(\phi^{(5)}+2\phi^{(3)}+\phi'\big)v\ n_1\ ds - R \\ &= -2(1-\nu)\big(\phi^{(5)}+2\phi^{(3)}+\phi', v_{,1}\big) - R \\ &= -2(1-\nu)\int_{\omega}\big(\phi^{(6)}+2\phi^{(4)}+\phi''\big)u\ d\alpha_2\ \partial_1(\alpha+1) - R + r \\ &= -4(1-\nu)\big((\phi^{(6)}+2\phi^{(4)}+\phi'', u)\big) - R + r. \end{aligned}$$

Finally,

$$\begin{aligned} V &= \nu\big(\alpha_1(\phi^{(4)}+\phi'')+\vartheta^{(3)}+\vartheta', w_{,11}\big) \\ &= \nu\big(\alpha_1(\phi^{(4)}+\phi'')+\vartheta^{(3)}+\vartheta', \beta_{22,11}-v_{,112}\big) \\ &= \nu\big(\alpha_1(\phi^{(4)}+\phi'')+\vartheta^{(3)}+\vartheta', \beta_{22,11}\big)+\nu\mathcal{O}(t), \end{aligned}$$

where the residual estimate follows from (3.2) and Theorem 4.1. Furthermore, by (4.1b), (4.23), and (4.24),

$$\begin{aligned} &\nu\big(\alpha_1(\phi^{(4)}+\phi'')+\vartheta^{(3)}+\vartheta', \beta_{22,11}\big) \\ &\quad = \nu\int_{\partial\omega}\big\{\big(\alpha_1(\phi^{(4)}+\phi'')+\vartheta^{(3)}+\vartheta'\big)\beta_{22,1} - \big(\phi^{(4)}+\phi''\big)\beta_{22}\big\}\ n_1\ ds = \nu\mathcal{O}(t^{1/2}). \end{aligned}$$

Combining these results and recalling (3.3) we have

$$\begin{aligned} &\mathcal{A}^K(\underline{u}^0, \underline{v}) + t^{-2}\mathcal{B}^K(\underline{u}^0, \underline{v}) \\ &\quad = q(\underline{v}) - 2(1-\nu)\int_{\partial\omega}\big(\phi^{(4)}+\phi''\big)\beta_{22}(\underline{v})\ n_1\ ds + \sum_{i,j=1}^{2}\big((a_{ij}, \beta_{ij}(\underline{v})\big) + \nu\mathcal{O}(t^{1/2}). \end{aligned}$$

Hence, by Theorem 2.1 and (4.23),

$$|||\underline{u}^K - \underline{u}^0|||_{K,t}^2 = \mathcal{O}(t + \nu t^{1/2}),$$

from which the assertion follows. □

COROLLARY 5.1. *Theorem* 4.1 *holds with* $\eta = t^{1/4}$ *if* $\|\cdot\| = \|\cdot\|_{L_2(\omega)}$ *and* $\eta = 1$ *if* $\|\cdot\| = \|\cdot\|_{L_\infty(\omega)}$.

Proof. Theorem 4.1 was proved by taking into account $\|\beta_{ij}^K\|_{L_2(\omega)} = \mathcal{O}(t)$ and $\|\kappa_{11}^K\|_{L_2(\omega)} = \mathcal{O}(1)$. However, $\|\beta_{ij}^K\|_{L_2(\omega)} = \mathcal{O}(t^{5/4})$ and $\|\kappa_{11}^K\|_{L_2(\omega)} = \mathcal{O}(t^{1/4})$. The improved estimates now follow by using same argument as in the proof of Theorem 4.1. □

Proof of (1.1c). Let $\underline{U} = \underline{U}^K - \underline{U}^0$ and $\underline{u} = \underline{u}^K - \underline{u}^0$. Then

$$
\begin{aligned}
e_{ij}(\underline{U}) &= \chi^{1/2}\{\beta_{ij}(\underline{u}) - \alpha_3\kappa(\underline{u}) + R_{ij}\}, \qquad i,j = 1,2,\\
e_{33}(\underline{U}) &= \chi^{1/2}\left\{-\frac{\nu}{1-\nu}\mathrm{tr}\underline{\underline{\beta}}(\underline{u}) + \alpha_3\frac{\nu}{1-\nu}\mathrm{tr}\underline{\underline{\kappa}}(\underline{u}) + R_{33}\right\},\\
e_{i3}(\underline{U}) &= \chi^{1/2}\cdot R_{i3}, \qquad i = 1,2,
\end{aligned}
\tag{5.1}
$$

where, by Theorem 5.1 and Corollary 5.1,

$$
\|R_{ij}\|_{L_2(\Omega)} = \begin{cases} \mathcal{O}(t^{9/4}) & \text{if } (i,j) = (1,3) \text{ or } (3,1),\\ \mathcal{O}(t^{11/4}) & \text{otherwise.}\end{cases}
$$

Applying Theorem 5.1 together with this estimate, we have

$$
\mathcal{A}^{3D}(\underline{U},\underline{U}) = \mathcal{A}^K(\underline{u},\underline{u}) + t^{-2}\mathcal{B}^K(\underline{u},\underline{u}) + \mathcal{O}(t^{3/2}). \tag{5.2}
$$

The first part of the assertion now follows from (5.2) and Theorem 5.1. Furthermore, in the Appendix we prove that the result of Theorem 5.1 cannot be improved in general. This, together with (5.2), implies the optimality of the given convergence rate. □

5.2. Proof of (1.1d): The soft membrane case. We are not only going to prove (1.1d), but we will also show that the given result is optimal in the general case. Let us begin with a basic theorem that comes out of analysis very similar to that in §4. Estimate (1.1d) will be an easy consequence of this result.

THEOREM 5.2. $|||\underline{u}^K - \underline{u}^M|||_{K,t} = \mathcal{O}(t^{5/4})$.

Proof. Without loss of generality we may consider a single Fourier component, so assume again that $f(\alpha_1,\alpha_2) = f^k(\alpha_1)\cos k\alpha_2$, $k \in \{0,1,2,\ldots\}$. It will be enough to handle separately cases where $k = 0$ and $k \geq 1$.

We begin with the axially symmetric case $k = 0$. By (4.3), $\underline{s}^t = \underline{u}^K - \underline{u}^M$ is of the form $(u_t(\alpha_1), w_t(\alpha_1))$ and

$$
\begin{aligned}
|||\underline{s}^t|||^2_{K,t} &= -t^2\mathcal{A}^F_0(\underline{u}^M,\underline{s}^t) = -t^2\int_I w''_M w''_t\, d\alpha_1\\
&= -t^2\left\{\Big/_I (w''_M w'_t - w^{(3)}_M w_t) + \int_I w^{(4)}_M \beta^t_{22}\, d\alpha_1\right\},
\end{aligned}
\tag{5.3}
$$

where $/_I g$ is the abbreviation for $g(1) - g(-1)$. Observe that the last equality above is the result of partial integration and the fact that $w_t = \beta_{22}(\underline{s}^t)$. Moreover,

$$
\|w_t\|_{0,I} \leq |||\underline{s}^t|||_{K,t}, \quad \|w_t\|_{2,I} \leq t^{-1}|||\underline{s}^t|||_{K,t}, \quad \|w_t\|_{1,I} \leq t^{-1/2}|||\underline{s}^t|||_{K,t},
$$

where the first two estimates follow from energy arguments and the last follows from from interpolation. The boundary values can be estimated by applying (4.7), with g and ϵ replaced by w_t and $t^{1/2}$, respectively. Accordingly, the right-hand side of (5.3) has an upper bound $\mathcal{O}(t^{5/4})|||\underline{s}^t|||_{k,t}$, and so the first case is proved.

Then assume that $k \geq 1$. By (4.3),

$$
\underline{s}^t = \big(u_t(\alpha_1)\cos k\alpha_2,\ v_t(\alpha_1)\sin k\alpha_2,\ w_t(\alpha_1)\cos k\alpha_2\big).
$$

Furthermore, $\underline{u}^t = (u_t, v_t, w_t) \in \mathcal{U}_k^F$ satisfies the variational equation

$$t^2\mathcal{A}_k^F(\underline{u}^t, \underline{v}) + \mathcal{B}_k^F(\underline{u}^t, \underline{v}) = -t^2\mathcal{A}_k^F(\underline{u}^M, \underline{v}), \quad \underline{v} \in H^1(I)^2 \times H^2(I). \tag{5.4}$$

Integrating by parts, we see that the right-hand side of (5.4) is equal to $q_1(\underline{v}) + q_2(\underline{v})$, where

$$q_1(\underline{v}) = t^2 \sum_{i,j} \int_I a_{ij}\beta_{ij}(\underline{v})\, d\alpha_1,$$

$$q_2(\underline{v}) = t^2 \Big/ \int_I \big(a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_{3,1}\big),$$

and where a_{ij} and a_i are smooth functions depending only on $\underline{u}^M$, thus independent on t. Accordingly, we can split $\underline{u}^t$ as $\underline{u}^1 + \underline{u}^2$, where $\underline{u}^i \in \mathcal{U}_k^F$ is the unique solution of

$$t^2\mathcal{A}_k^F(\underline{u}^i, \underline{v}) + \mathcal{B}_k^F(\underline{u}^i, \underline{v}) = q_i(\underline{v}), \quad \underline{v} \in H^1(I)^2 \times H^2(I), \ i = 1, 2.$$

Then, obviously, $|||\underline{u}^1|||_k = \mathcal{O}(t^2)$, so it suffices to consider $\underline{u}^2 = (u_2, v_2, w_2)$. First, by the energy argument

$$\|u_2\|_{1,I} + \|v_2\|_{1,I} + \|w_2\|_{2,I} = \mathcal{O}(1). \tag{5.5}$$

Furthermore, $\underline{u}^2$ satisfies the Euler equations (4.5) with $f^k = 0$ and the natural boundary conditions (4.6) with the right side replaced by $\mathcal{O}(t^2)$ in (4.6a) and (4.6b) and by $\mathcal{O}(1)$ in (4.6c) and (4.6d). Hence, proceeding as in §4, we conclude that (4.16) and (4.18) hold with w and w_0 replaced by w_2 and 0, respectively. Then, observing that u_2 and v_2 have the same fundamental set as w_2 (see (4.12) and (4.14)), it follows that $\underline{u}^2$ is of the form

$$\underline{u}^2 = \left(\sum_{i=0}^{7} U^i\Psi_i, \ \sum_{i=0}^{7} V^i\Psi_i, \ \sum_{i=0}^{7} W^i\Psi_i\right). \tag{5.6}$$

By (4.18), $W^i = \mathcal{O}(t)$ for $i = 4, 5, 6, 7$, and by (5.5), all the remaining coefficients are of order $\mathcal{O}(1)$ at most. Upon substituting (5.6) into the Euler equations (4.5) (with $f^k = 0$), we can solve U^i and V^i in terms of W^0, W^1, W^2, and W^3, when $i = 0, 1, 2, 3$, and in terms of W^4, W^5, W^6, and W^7, when $i = 4, 5, 6, 7$. Applying (4.13) and (4.14), we get

$$\begin{aligned} U^0 &= -\frac{1}{k^2}W^1 - \frac{6(2+\nu)}{k^4}W^3 + \mathcal{O}(t^2), \\ U^1 &= -\frac{2}{k^2}W^2 + \mathcal{O}(t^2), \\ U^2 &= -\frac{3}{k^2}W^3 + \mathcal{O}(t^2), \\ U^3 &= \mathcal{O}(t^2), \end{aligned} \tag{5.7a}$$

$$\begin{aligned} V^0 &= -\frac{1}{k}W^0 + \frac{2\nu}{k^3}W^2 + \mathcal{O}(t^2), \\ V^1 &= -\frac{1}{k}W^1 + \frac{6\nu}{k^3}W^3 + \mathcal{O}(t^2), \\ V^2 &= -\frac{1}{k}W^2 + \mathcal{O}(t^2), \\ V^3 &= -\frac{1}{k}W^3 + \mathcal{O}(t^2), \end{aligned} \tag{5.7b}$$

$$
\begin{aligned}
U^4 &= \frac{\nu}{2A_1}W^5 - \frac{\nu}{2A_1}W^7 + \mathcal{O}(t^{5/2}) = \mathcal{O}(t^{3/2}),\\
U^5 &= -\frac{\nu}{2A_1}W^4 - \frac{\nu}{2A_1}W^6 + \mathcal{O}(t^{5/2}) = \mathcal{O}(t^{3/2}),\\
U^6 &= -\frac{\nu}{2A_1}W^5 - \frac{\nu}{2A_1}W^7 + \mathcal{O}(t^{5/2}) = \mathcal{O}(t^{3/2}),\\
U^7 &= -\frac{\nu}{2A_1}W^4 + \frac{\nu}{2A_1}W^6 + \mathcal{O}(t^{5/2}) = \mathcal{O}(t^{3/2}),
\end{aligned}
\tag{5.7c}
$$

$$
\begin{aligned}
V^4 &= -\frac{k(2+\nu)}{2A_1^2}W^6 + \mathcal{O}(t^3) = \mathcal{O}(t^2),\\
V^5 &= \frac{k(2+\nu)}{2A_1^2}W^7 + \mathcal{O}(t^3) = \mathcal{O}(t^2),\\
V^6 &= \frac{k(2+\nu)}{2A_1^2}W^4 + \mathcal{O}(t^3) = \mathcal{O}(t^2),\\
V^7 &= -\frac{k(2+\nu)}{2A_1^2}W^5 + \mathcal{O}(t^3) = \mathcal{O}(t^2).
\end{aligned}
\tag{5.7d}
$$

Here the residual estimates are based on the above-mentioned upper bounds for the coefficients, and $A_1 = \mathcal{O}(t^{-1/2})$ is defined as in (4.9). Consequently,

$$
|||\underline{u}^2|||_k \le C(k)\big\{t(k-1)\big(|W^0| + |W^1|\big) + |W^2| + |W^3| + t^{5/4}\big\}. \tag{5.8}
$$

Applying (5.6), (5.7), (4.12), (4.13), (4.14), and the natural boundary conditions (4.6a) and (4.6b) (with the right side replaced by $\mathcal{O}(t^2)$), we have, by straightforward computation, that $W^i = \mathcal{O}(t^2)$, $i = 2,3$. Hence, it remains to show that $W^i = \mathcal{O}(t^{1/4})$, $i = 0,1$ (if $k \ge 2$). In fact, we obtain a better estimate. First recalling that $\mathcal{A}^K(\underline{s}^t, \underline{v}) = 0$ for all $\underline{v} \in \boldsymbol{\mathcal{U}}^0$, we conclude that

$$
\begin{aligned}
S_1(\underline{u}^t) &:= \int_I \big(k\alpha_1(\nu\kappa_{11}(\underline{u}^t) + \kappa_{22}(\underline{u}^t)) + 2(1-\nu)\kappa_{12}(\underline{u}^t)\big)\, d\alpha_1 = 0,\\
S_2(\underline{u}^t) &:= \int_I \big(\nu\kappa_{11}(\underline{u}^t) + \kappa_{22}(\underline{u}^t)\big)\, d\alpha_1 = 0,
\end{aligned}
$$

and so by the Cauchy–Schwarz inequality, $|S_i(\underline{u}^2)| = |S_i(\underline{u}^1)| = \mathcal{O}(t)$, $i = 1,2$. But recalling the already-known estimates, we have

$$
\begin{aligned}
S_1(\underline{u}^2) &= -\left(\frac{2}{3}(k^3-k)\nu + 4(1-\nu)\left(k - \frac{1}{k}\right)\right)W^1 + \mathcal{O}(t + \nu t^{1/2}),\\
S_2(\underline{u}^1) &= -2(k^2-1)W^0 + \mathcal{O}(t^{3/2} + \nu t^{1/2}).
\end{aligned}
$$

So $W^i = \mathcal{O}(t + \nu t^{1/2})$, $i = 0,1$, and the proof of Theorem 5.2 is thus complete. □

Remark 5.1. Let $\omega_\delta = \{\underline{\alpha} \in \omega \mid -1+\delta < \alpha_1 < 1-\delta\}$, $0 < \delta < 1$, and let $\mathcal{A}_\delta^K(\underline{u}, \underline{u})$, $\mathcal{B}_\delta^K(\underline{u}, \underline{u})$, and $|||\cdot|||_{K,t,\delta}$ be the corresponding interior bending and membrane strain energies and the interior energy norm, respectively. Then it follows from the above analysis (also carried further in the case $k = 0$) that

$$
|||\underline{u}^K - \underline{u}^M|||_{K,t,\delta} = \mathcal{O}(t^2 + \nu t^{3/2}), \qquad \delta \ge t^{1/2}\ln(t^{-3/4}),
$$

so the interior accuracy of the asymptotic model is better than the global one.

Proof of (1.1d). Let $\underline{U} = \underline{U}^K - \underline{U}^M$ and $\underline{u} = \underline{u}^K - \underline{u}^M = (u, v, w)$. Recalling Theorem 5.2 and proceeding as in the proof of Theorem 4.1, we have

$$\begin{aligned} \|D^\tau u\|_{L_2(\omega)} &= \mathcal{O}(t^{1/4} + t^{(7-2\tau_1)/4}), \\ \|D^\tau v\|_{L_2(\omega)} &= \mathcal{O}(t^{1/4} + t^{(9-2\tau_1)/4}), \\ \|D^\tau w\|_{L_2(\omega)} &= \mathcal{O}(t^{1/4} + t^{(5-2\tau_1)/4}), \end{aligned} \tag{5.9}$$

for every multi-index $\tau = (\tau_1, \tau_2)$. Applying (5.9), we have, as in the proof of (1.1c), that

$$|||\underline{U}|||_{3D}^2 = |||\underline{u}|||_{K,t}^2 + \mathcal{O}(t^3), \tag{5.10}$$

from which (1.1d) follows.

To prove the optimality of (1.1d), let $f = \alpha_1^2$. By a straightforward computation,

$$\begin{aligned} u(\alpha_1) &= -\frac{\nu(W^4 + W^6)}{2A}\Psi_5(\alpha_1) - \frac{\nu(W^4 - W^6)}{2A}\Psi_7(\alpha_1), \\ w(\alpha_1) &= W^4\Psi_4(\alpha_1) + W^6\Psi_6(\alpha_1), \end{aligned}$$

where $A = \big(3(1-\nu^2)\big)^{1/4} t^{-1/2}$,

$$W^4 = -\frac{\cos A - \sin A}{12A^2(1-\nu^2)} + \mathcal{O}(A^{-2}e^{-2A}), \quad W^6 = -\frac{\cos A + \sin A}{12A^2(1-\nu^2)} + \mathcal{O}(A^{-2}e^{-2A}),$$

and Ψ_4 through Ψ_7 are defined as in (4.15). It is easy to check that $|||\underline{u}|||_{K,t} \sim t^{5/4}$, and thus by (5.10), the asserted convergence rate is optimal. □

6. Convergence rate estimate (1.1b). In the previous section we had to analyze the edge behaviour of $\underline{u}^K$ in quite a bit of detail before we could prove the basic Theorems 5.1 and 5.2. The corresponding estimates for $\underline{u}^R - \underline{u}^K$ in case of both deformation states, which easily imply (1.1b), can be obtained without any such special information about $\underline{u}^R$. However, this time we cannot guarantee the optimality of the convergence rate.

We begin with the following lemma.

LEMMA 6.1. $\big|\int_{\partial\omega} \kappa_{12}^K \rho_2^R n_1 \, ds\big| \le C(f)\sigma^{-1}t^{1/2} \cdot |||\underline{u}^R|||_{R,t}$, *where* σ *is defined by* (1.2).

Proof. Assume that $f = f^k(\alpha_1)\cos k\alpha_2$, $k \ge 1$. (If $k = 0$, the left side vanishes.) In that case

$$\underline{u}^R = \big(u \cdot \cos k\alpha_2,\ v \cdot \sin k\alpha_2,\ w \cdot \cos k\alpha_2,\ \theta_1 \cdot \cos k\alpha_2,\ \theta_2 \cdot \sin k\alpha_2\big),$$

where

$$\underline{u} = (u, v, w, \theta_1, \theta_2) \in \left\{ H^1(I)^5 \mid \int_I (v - w) = \int_I \alpha_1(v - w) = 0 \text{ if } k = 1 \right\}$$

is such that

$$\sigma^2\big\{\mathcal{A}_k^F(\underline{u}, \underline{\tilde{v}}) + t^{-2}\mathcal{B}_k^F(\underline{u}, \underline{\tilde{v}}) + t^{-2}\mathcal{C}_k^F(\underline{u}, \underline{\tilde{v}})\big\} = \int_I f^k \cdot \tilde{w} \, d\alpha_1, \quad \underline{\tilde{v}} \in H^1(I)^5. \tag{6.1}$$

Here, $\mathcal{A}_k^F(\underline{u},\underline{\tilde{v}})$, $\mathcal{B}_k^F(\underline{u},\underline{\tilde{v}})$, and $\mathcal{C}_k^F(\underline{u},\underline{\tilde{v}})$ are otherwise of the form (2.4), (2.5), and (2.6), respectively, but now ω and $d\underline{\alpha}$ are replaced by I and $d\alpha_1$, $\underline{\underline{\beta}}$ is as in (4.4), and, finally,

$$
\begin{aligned}
&\kappa_{11} = \theta_1', && \\
&\kappa_{12} = \frac{1}{2}\,(-k\theta_1 + \theta_2' - v'), && \rho_1 = -\theta_1 + w', \\
&\kappa_{22} = k\theta_2, && \rho_2 = -(\theta_2 + kw + v).
\end{aligned} \tag{6.2}
$$

Then, $\kappa_{12}^K(\alpha_1,\alpha_2) = \kappa_{12}(\alpha_1)\sin k\alpha_2$, $\rho_2^R(\alpha_1,\alpha_2) = \rho_2(\alpha_1)\sin k\alpha_2$, and thus

$$
\left| \int_{\partial\omega} \kappa_{12}^K\, \rho_2^R\, n_1\, ds \right| \le \pi \cdot \left(|\kappa_{12}(1)\rho_2(1)| + |\kappa_{12}(-1)\rho_2(-1)| \right).
$$

By the energy argument and by (6.2) and (4.7), with g and ϵ replaced by ρ_2 and t, respectively,

$$
|\rho_2(\pm 1)| \le C\left\{ t^{-1/2}\|\rho_2\|_{L_2(I)} + t^{1/2}\|\rho_2'\|_{L_2(I)} \right\} \le Ck\sigma^{-1}t^{1/2}|||\underline{u}^R|||_{R,t}.
$$

Furthermore, by Theorem 4.1 and (5.9), $|\kappa_{12}(\pm 1)| = \mathcal{O}(1)$. Combining these estimates, the assertion follows. □

In the remaining part of this section we use the exceptional notation

$$
\underline{u}^K = (u^K,\ v^K,\ w^K,\ w,_1^K,\ w,_2^K - v^K).
$$

THEOREM 6.1. $|||\underline{u}^R - \underline{u}^K|||_{R,t} = \mathcal{O}(\sigma t^{1/2})$, *where* σ *is defined by* (1.2).

Proof. We note first that (u^K, v^K, w^K) satisfies the Euler equations

$$
\beta_{11,1}^K + (1-\nu)\beta_{12,2}^K + \nu\beta_{22,1}^K = 0, \tag{6.3a}
$$

$$
12t^{-2}\nu\beta_{11,2}^K + 12t^{-2}(1-\nu)\beta_{12,1}^K + 12t^{-2}\beta_{22,2}^K - K_1 = 0, \tag{6.3b}
$$

$$
12t^{-2}\nu\beta_{11}^K + 12t^{-2}\beta_{22}^K + K_2 = \sigma^{-2}f, \tag{6.3c}
$$

where

$$
\begin{aligned}
K_1 &= \nu\kappa_{11,2}^K + 2(1-\nu)\kappa_{12,1}^K + \kappa_{22,2}^K, \\
K_2 &= \kappa_{11,11}^K + \nu\kappa_{22,11}^K + 2(1-\nu)\kappa_{12,12}^K + \nu\kappa_{11,22}^K + \kappa_{22,22}^K,
\end{aligned}
$$

and the natural boundary conditions

$$
\beta_{11}^K + \nu\beta_{22}^K = 0, \tag{6.4a}
$$

$$
6t^{-2}\beta_{12}^K - \kappa_{12}^K = 0, \tag{6.4b}
$$

$$
\kappa_{11}^K + \nu\kappa_{22}^K = 0, \tag{6.4c}
$$

$$
\kappa_{11,1}^K + \nu\kappa_{22,1}^K + 2(1-\nu)\kappa_{12,2}^K = 0. \tag{6.4d}
$$

Let

$$
\begin{aligned}
&\underline{v}_1 = (u,0,0,0,0), && \underline{v}_2 = (0,v,0,0,0), && \underline{v}_3 = (0,0,w,0,0), \\
&\underline{v}_4 = (0,0,0,\theta_1,0), && \underline{v}_5 = (0,0,0,0,\theta_2),
\end{aligned}
$$

where $\underline{v} = (u,v,w,\theta_1,\theta_2) = \underline{u}^R - \underline{u}^K$; let us use the abbreviation $a^R(\underline{u},\underline{v})$ for the Reissner–Mindlin inner product on the left side of (2.13) and the abbreviation $\langle g,h\rangle$

for the boundary integral $\int_{\partial\omega} g\cdot h\, n_1\, ds$. Then it follows from (6.3a), (6.4a) and (6.4b) that

$$\sigma^{-2}a^R(\underline{u}^K,\underline{v}_1)=0, \tag{6.5}$$

$$\sigma^{-2}a^R(\underline{u}^K,\underline{v}_2)=-\big(\nu\kappa^K_{11,2}+\kappa^K_{22,2}+(1-\nu)\kappa^K_{12,1},v\big)+(1-\nu)\langle\kappa^K_{12},v\rangle. \tag{6.6}$$

Furthermore, we have

$$\sigma^{-2}a^R(\underline{u}^K,\underline{v}_3)=\big(12t^{-2}\nu\beta^K_{11}+12t^{-2}\beta^K_{22},w\big), \tag{6.7}$$

and by (6.3c) and (6.4d),

$$\begin{aligned}\sigma^{-2}a^R(\underline{u}^K,\underline{v}_4)&=\big(-\kappa^K_{11,1}-\nu\kappa^K_{22,1}-(1-\nu)\kappa^K_{12,2},\theta_1\big)\\&=\big(\kappa^K_{11,11}+\nu\kappa^K_{22,11}+(1-\nu)\kappa^K_{12,12},w\big)+\big(\kappa^K_{11,1}+\nu\kappa^K_{22,1}+(1-\nu)\kappa^K_{12,2},\rho^R_1\big)\\&\quad-\langle\kappa^K_{11,1}+\nu\kappa^K_{22,1}+(1-\nu)\kappa^K_{12,2},w\rangle\\&=\big(\kappa^K_{11,11}+\nu\kappa^K_{22,11}+(1-\nu)\kappa^K_{12,12},w\big)+\big(\kappa^K_{11,1}+\nu\kappa^K_{22,1}+(1-\nu)\kappa^K_{12,2},\rho^R_1\big)\\&\quad-(1-\nu)\langle\kappa^K_{12},w_{,2}\rangle.\end{aligned} \tag{6.8}$$

Finally,

$$\begin{aligned}\sigma^{-2}a^R(\underline{u}^K,\underline{v}_5)&=-\big(\nu\kappa^K_{11,2}+\kappa^K_{22,2}+(1-\nu)\kappa^K_{12,1},\theta_2\big)+(1-\nu)\langle\kappa^K_{12},\theta_2\rangle\\&=\big(\nu\kappa^K_{11,22}+\kappa^K_{22,22}+(1-\nu)\kappa^K_{12,12},w\big)+\big(\nu\kappa^K_{11,2}+\kappa^K_{22,2}+(1-\nu)\kappa^K_{12,1},v\big)\\&\quad+\big(\nu\kappa^K_{11,2}+\kappa^K_{22,2}+(1-\nu)\kappa^K_{12,1},\rho^R_2\big)+(1-\nu)\langle\kappa^K_{12},\theta_2\rangle.\end{aligned} \tag{6.9}$$

Combining (6.5)–(6.9) and applying (6.3c), we have

$$\begin{aligned}|||\underline{v}|||^2_{R,t}&=-\sigma^2\big(\kappa^K_{11,1}+\nu\kappa^K_{22,1}+(1-\nu)\kappa^K_{12,2},\rho^R_1\big)\\&\quad-\sigma^2\big(\nu\kappa^K_{11,2}+\kappa^K_{22,2}+(1-\nu)\kappa^K_{12,1},\rho^R_2\big)+(1-\nu)\sigma^2\langle\kappa^K_{12},\rho^R_2\rangle.\end{aligned}$$

The assertion now follows from Lemma 6.1, Corollary 5.1, and (5.9). □

COROLLARY 6.1. *Both in bending and soft membrane cases,* $\underline{u}^R\in[C^\infty_{\rm per}(\overline{\omega})]^5$ *and*

$$\begin{aligned}&\|\beta^R_{ij}\|_{L_2(\omega)}=\mathcal{O}(\sigma^{-5/4}t^{5/4}),\\&\|\rho^R_i\|_{L_2(\omega)}=\mathcal{O}(t^{3/2}),\\&\|(u^R,v^R,w^R)\|_{m,\omega}=\mathcal{O}(1),\quad m=1,2,\\&\|(\theta^R_1,\theta^R_2)\|_{m,\omega}=\mathcal{O}(1+t^{(1-m)/2}),\quad m=1,2.\end{aligned}$$

Proof. Assume again that $f=f(\alpha_1)\cos k\alpha_2$. In that case, the argument proceeds as in the proof of Lemma 6.1, and thus the first two estimates and also the remaining estimates with $m=1$ follow immediately from Theorems 5.1 and 5.2, (5.9), and Theorem 6.1. Then, proceeding from the Euler equations corresponding to (6.1), applying the already-obtained estimates, and continuing as in the proof of Theorem 4.1, the smoothness result as well as the remaining estimates follow. □

Proof of (1.1b). Applying Corollary 6.1 and proceeding as in the proof of Theorem (1.1c), the assertion follows easily. □

7. Proof of (1.1a).

7.1. Proof of (1.1a): The bending-dominated case. We obtained results (1.1b), (1.1c), and (1.1d) by using straightforward energy arguments. In the proof of (1.1a), the complementary energy principle is used instead. This is necessary mainly because we do not know the t-dependence of constant $c(t)$ in (2.10). The complementary energy principle was first applied in the plate theory by D. Morgenstern in 1959 [M], and by many authors afterwards (see, e.g., [K]). The statically admissible stress tensor given below is constructed in the spirit of [K]. However, we avoid Kichhoff-type hypotheses and instead directly use the Euler equations that $\underline{u}^R$ satisfies. These are

$$\beta^R_{11,1} + (1-\nu)\beta^R_{12,2} + \nu\beta^R_{22,1} = 0, \tag{7.1a}$$

$$12\nu\beta^R_{11,2} + 12(1-\nu)\beta^R_{12,1} + 12\beta^R_{22,2} + 6(1-\nu)\rho^R_2 - t^2(1-\nu)\kappa^R_{12,1} = 0, \tag{7.1b}$$

$$12t^{-2}\nu\beta^R_{11} + 12t^{-2}\beta^R_{22} - 6t^{-2}(1-\nu)\big(\rho^R_{1,1} + \rho^R_{2,2}\big) = f, \tag{7.1c}$$

$$6t^{-2}(1-\nu)\rho^R_1 + \kappa^R_{11,1} + (1-\nu)\kappa^R_{12,2} + \nu\kappa^R_{22,1} = 0, \tag{7.1d}$$

$$6t^{-2}(1-\nu)\rho^R_2 + \nu\kappa^R_{11,2} + (1-\nu)\kappa^R_{12,1} + \kappa^R_{22,2} = 0 \tag{7.1e}$$

with the following natural boundary conditions at both ends:

$$\beta^R_{11} + \nu\beta^R_{22} = 0, \quad \beta^R_{12} = 0, \quad \rho^R_1 = 0, \tag{7.2a}$$

$$\kappa^R_{11} + \nu\kappa^R_{22} = 0, \quad \kappa^R_{12} = 0. \tag{7.2b}$$

Let $\mathcal{H}$ stand for the space of stress tensor defined as

$$\mathcal{H} = \big\{\underline{\underline{s}} = (s_{ij})_{i,j=1}^3 \mid s_{ij} \in L_2(\Omega),\ s_{ij} = s_{ji}\big\},$$

and let $S : \mathcal{H} \to \mathcal{H}$ stand for the isomorphism defined by

$$\big(S\underline{\underline{\tau}}\big)_{ij} = D^{-1}\big(\lambda \mathrm{tr}\underline{\underline{\tau}}\delta_{ij} + \mu\tau_{ij}\big), \quad \big(S^{-1}\underline{\underline{\tau}}\big)_{ij} = \frac{D}{E}\big(-\nu \mathrm{tr}\underline{\underline{\tau}}\delta_{ij} + (1+\nu)\tau_{ij}\big),$$

where D, E, μ, λ, and ν are defined as in §2 and δ_{ij} is the Kronecker symbol. Then S and S^{-1} are obviously self-adjoint if $\mathcal{H}$ is supplied with the weighted L_2 inner product

$$(\underline{\underline{\eta}}, \underline{\underline{\tau}})_{\mathcal{H}} = \sum_{i,j=1}^3 \int_\Omega \eta_{ij}\tau_{ij}(1+\alpha_3)\, d\underline{\alpha}.$$

Using this notation we have $\mathcal{A}^{3D}(\underline{U}, \underline{V}) = \big(S\underline{\underline{e}}(\underline{U}), \underline{\underline{e}}(\underline{V})\big)_{\mathcal{H}}$.

Next, let $\underline{\underline{s}}^{3D} = S\underline{\underline{e}}(\underline{U}^{3D})$ and define the set

$$\mathcal{H}_Q = \big\{\underline{\underline{s}} \in \mathcal{H} \mid \big(\underline{\underline{s}}, \underline{\underline{e}}(\underline{V})\big)_{\mathcal{H}} = Q^{3D}(\underline{V}),\ \underline{V} \in \mathcal{W}^{3D}\big\}.$$

This is usually referred to as the set of statically admissible stresses. We note that $\underline{\underline{s}}^{3D} \in \mathcal{H}_Q$. Let us further define in $\mathcal{H} \times \mathcal{H}$ the bilinear form $\mathcal{B}^{3D}$ and define in $\mathcal{H}$ the "complementary" energy G^{3D} as

$$\mathcal{B}^{3D}(\underline{\underline{\eta}}, \underline{\underline{\tau}}) = \big(S^{-1}\underline{\underline{\eta}}, \underline{\underline{\tau}}\big)_{\mathcal{H}}, \qquad G^{3D}(\underline{\underline{\eta}}) = \frac{1}{2}\mathcal{B}^{3D}(\underline{\underline{\eta}}, \underline{\underline{\eta}}).$$

The key argument in the proof of (1.1a) is the following lemma, which may be viewed as a version of the well-known complementary energy principle. Sometimes this is also referred to as the "hypercircle theorem" [M].

LEMMA 7.1. *For any* $\underline{U} \in \mathcal{W}^{3D}$ *and* $\underline{\underline{s}} \in \mathcal{H}_Q$,

$$\frac{1}{2}\mathcal{A}^{3D}(\underline{U}^{3D} - \underline{U}, \underline{U}^{3D} - \underline{U}) + \frac{1}{2}\mathcal{B}^{3D}(\underline{\underline{s}}^{3D} - \underline{\underline{s}}, \underline{\underline{s}}^{3D} - \underline{\underline{s}}) = F^{3D}(\underline{U}) + G^{3D}(\underline{\underline{s}}).$$

Proof. Rewrite the left side of the asserted identity as

$$\frac{1}{2}\{\mathcal{A}^{3D}(\underline{U}^{3D}, \underline{U}^{3D}) + \mathcal{B}^{3D}(\underline{\underline{s}}^{3D}, \underline{\underline{s}}^{3D})\} + \left\{\frac{1}{2}\mathcal{A}^{3D}(\underline{U}, \underline{U}) - \mathcal{A}^{3D}(\underline{U}^{3D}, \underline{U})\right\}$$
$$+ \frac{1}{2}\mathcal{B}^{3D}(\underline{\underline{s}}, \underline{\underline{s}}) - \mathcal{B}^{3D}(\underline{\underline{s}}^{3D}, \underline{\underline{s}}) = Q^{3D}(\underline{U}^{3D}) + F^{3D}(\underline{U}) + G^{3D}(\underline{\underline{s}}) - \mathcal{B}^{3D}(\underline{\underline{s}}^{3D}, \underline{\underline{s}}),$$

and note that $\mathcal{B}^{3D}(\underline{\underline{s}}^{3D}, \underline{\underline{s}}) = \big(\underline{\underline{s}}, \underline{\underline{e}}(\underline{U}^{3D})\big)_{\mathcal{H}} = Q^{3D}(\underline{U}^{3D})$ for any $\underline{\underline{s}} \in \mathcal{H}_Q$. □

Hence, our goal is to find $\underline{\underline{s}} \in \mathcal{H}_Q$ such that

$$F^{3D}(\underline{U}^R) + G^{3D}(\underline{\underline{s}}) = \mathcal{O}(t).$$

The assertion then follows from Lemma 7.1. Now, $\underline{\underline{s}} \in \mathcal{H}_Q$ if $\underline{\underline{s}} \in H^1(\Omega)^{3\times 3}$ satisfies the Euler equations

$$\begin{aligned} &\chi^{-1}s_{11,1} + s_{12,2} + (\chi^{-1}s_{13}),_3 = 0,\\ &\chi^{-1}s_{12,1} + s_{22,2} + \chi(\chi^{-2}s_{23}),_3 = 0,\\ &\chi^{-1}s_{13,1} + s_{23,2} + (\chi^{-1}s_{33}),_3 - s_{22} = 0, \end{aligned} \tag{7.3}$$

and the natural boundary conditions

$$\begin{aligned} &(s_{11}, s_{12}, s_{13})(\pm 1, \alpha_2, \alpha_3) = 0, && (s_{13}, s_{23})\left(\alpha_1, \alpha_2, \pm\frac{t}{2}\right) = 0,\\ &s_{33}\left(\alpha_1, \alpha_2, \frac{t}{2}\right) = f(\alpha_1, \alpha_2), && s_{33}\left(\alpha_1, \alpha_2, -\frac{t}{2}\right) = 0. \end{aligned} \tag{7.4}$$

We choose

$$\begin{aligned} s_{11} &= C\cdot\chi\{(\beta_{11}^R + \nu\beta_{22}^R) - \alpha_3(\kappa_{11}^R + \nu\kappa_{22}^R)\},\\ s_{12} &= C\cdot\{(1-\nu)\beta_{12}^R - \alpha_3(1-\nu)\kappa_{12}^R\},\\ s_{13} &= C\cdot -\chi\left(\frac{\alpha_3^2}{2} - \frac{t^2}{8}\right)6t^{-2}(1-\nu)\rho_1^R,\\ s_{22} &= C\cdot\{\nu\beta_{11}^R + \beta_{22}^R + R_{21} + R_{22} + R_{23}\},\\ s_{23} &= C\cdot -\left(\frac{\alpha_3^2}{2} - \frac{t^2}{8}\right)6t^{-2}(1-\nu)\rho_2^R,\\ s_{33} &= C\cdot\chi\Big\{\left(\frac{\alpha_3^3}{6} - \frac{\alpha_3 t^2}{8} - \frac{t^3}{24}\right)6t^{-2}(1-\nu)(\rho_{1,1}^R + \rho_{2,2}^R)\\ &\qquad + \left(\alpha_3 + \frac{t}{2}\right)(\nu\beta_{11}^R + \beta_{22}^R) + R_{31} + R_{32} + R_{33}\Big\}, \end{aligned}$$

where $C = 12(1 + t/2)t^{-3}$ and, furthermore,

$$\begin{aligned} R_{31} &= -\left(\frac{\alpha_3^2}{2} - \frac{t^2}{8}\right)(1-\nu)\int_0^{\alpha_2}\beta_{12,1}^R(\alpha_1, x)\,dx,\\ R_{32} &= -\left(\frac{\alpha_3^2}{2} - \frac{t^2}{8} + \frac{2\alpha_3^3}{3} - \frac{\alpha_3 t^2}{6}\right)(\nu\kappa_{11}^R + \kappa_{22}^R),\\ R_{33} &= -\left(\frac{\alpha_3^3}{3} - \frac{\alpha_3 t^2}{12}\right)(1-\nu)\int_0^{\alpha_2}\kappa_{12,1}^R(\alpha_1, x)\,dx,\\ R_{2i} &= R_{3i,3}, \quad i = 1,2,3. \end{aligned}$$

It may be checked from (7.1) and (7.2) that $\underline{\underline{s}}$ satisfies (7.3) and (7.4), and thus $\underline{\underline{s}} \in \mathcal{H}_Q$. We can write $\underline{\underline{s}}$ in the more convenient form

$$\begin{aligned}
s_{11} &= \frac{12}{t^3}\big(\beta_{11}^R + \nu\beta_{22}^R - \alpha_3(\kappa_{11}^R + \nu\kappa_{22}^R)\big) + r_{11},\\
s_{12} &= \frac{12}{t^3}(1-\nu)\big(\beta_{12}^R - \alpha_3\kappa_{12}^R\big) + r_{12},\\
s_{22} &= \frac{12}{t^3}\big(\nu\beta_{11}^R + \beta_{22}^R - \alpha_3(\nu\kappa_{11}^R + \kappa_{22}^R)\big) + r_{22},\\
s_{13} &= r_{13}, \qquad s_{23} = r_{23}, \qquad s_{33} = r_{33},
\end{aligned}$$

where, by Corollary 6.1, the residual terms are such that

$$G^{3D}(\underline{\underline{s}}) = \frac{1}{2}\big\{\mathcal{A}^R(\underline{u}^R,\underline{u}^R) + t^{-2}\mathcal{B}^R(\underline{u}^R,\underline{u}^R) + t^{-2}\mathcal{C}^R(\underline{u}^R,\underline{u}^R)\big\} + \mathcal{O}(t).$$

Furthermore, again recalling Corollary 6.1 we have

$$F^{3D}(\underline{U}^R) = \frac{1}{2}\big\{\mathcal{A}^R(\underline{u}^R,\underline{u}^R) + t^{-2}\mathcal{B}^R(\underline{u}^R,\underline{u}^R) + t^{-2}\mathcal{C}^R(\underline{u}^R,\underline{u}^R)\big\} - q(\underline{u}^R) + \mathcal{O}(t).$$

The assertion now follows recalling Lemma 7.1. □

7.2. Proof of (1.1a): The soft membrane case. We begin with the following definition. Let $\underline{\tilde{U}}^{3D} \in \mathcal{U}^{3D}$ be such that

$$\mathcal{A}(\underline{\tilde{U}}^{3D},\underline{V}) = \int_\omega f(\alpha_1,\alpha_2)V_3\left(\alpha_1,\alpha_2,\frac{t}{2}\right)\,d\underline{\alpha}, \qquad \underline{V} \in \mathcal{U}^{3D}.$$

By linearity, $\underline{U}^{3D} = (1+t/2)\underline{\tilde{U}}^{3D}$, and it follows from the triangle inequality, together with (1.1b), (1.1c), and (1.1d), that

$$|||\underline{U}^{3D} - \underline{U}^R|||_{3D} \le \left(1+\frac{t}{2}\right)|||\underline{\tilde{U}}^{3D} - \underline{U}^M|||_{3D} + \mathcal{O}(t).$$

Hence, it suffices to prove that $|||\underline{\tilde{U}}^{3D} - \underline{U}^M|||_{3D} = \mathcal{O}(t)$.

We first note that

$$\begin{aligned}
e_{11}(\underline{U}^M) &= \beta_{11} + R_{11}, & e_{22}(\underline{U}^M) &= \beta_{22} + R_{22},\\
e_{12}(\underline{U}^M) &= (1+\alpha_3)^{-1}\beta_{12} + R_{12}, & e_{23}(\underline{U}^M) &= R_{23},\\
e_{13}(\underline{U}^M) &= R_{13}, & e_{33}(\underline{U}^M) &= -\frac{\nu}{1-\nu}\mathrm{tr}\underline{\underline{\beta}} + R_{33},
\end{aligned}$$

where the remainder terms are of order

$$\|R_{ij}\|_{L_2(\Omega)} = \mathcal{O}(t^{3/2}). \tag{7.5}$$

Next, let $\underline{V} \in \mathcal{W}^{3D}$. Integrating by parts and applying (3.6b) and (3.6c), we have

$$\begin{aligned}
\mathcal{A}^{3D}(\underline{U}^M,\underline{V}) &= \frac{12}{t}\int_\Omega(\nu\beta_{11}+\beta_{22})V_3\,d\underline{\alpha} + R_I\\
&= \frac{12}{t}\int_\Omega(\nu\beta_{11}+\beta_{22})V_3\,d\alpha_1 d\alpha_2\partial_3\left(\alpha_3+\frac{t}{2}\right) + R_I\\
&= \int_\omega f(\alpha_1,\alpha_2)V_3\left(\alpha_1,\alpha_2,\frac{t}{2}\right)\,d\underline{\alpha} + R_I + R_{II},
\end{aligned}$$

where by (7.5), $R(\underline{V}) = R_I + R_{II}$ is such that $|R(\underline{V})| \leq Ct|||\underline{V}|||_{3D}$. The assertion then follows noting that $|||\tilde{\underline{U}}^{3D} - \underline{U}^M|||_{3D}^2 = -R(\tilde{\underline{U}}^{3D} - \underline{U}^M)$. □

Remark 7.1. Note that since we have dropped out the factor $1 + t/2$ from the potential energy in our shell model, there is no reason to expect faster convergence than that predicted by the above theorem.

Appendix. Here we analyze the behavior of $\underline{u}^K$ in more detail in two special cases where the deformation state is bending dominated. The load is chosen to be either (1) $f = K_0 \cos 2\alpha_2$ or (2) $f = K_1\alpha_1 \cos 2\alpha_2$, and we set $\nu = 1/3$ in both cases.

(1) Let $f = K_0 \cos 2\alpha_2$. In this case $\underline{u}^K$ is of the form (4.3), where $(u, v, w)(\alpha_1)$ satisfies (4.5) and (4.6). It is easy to check that

$$\tilde{u} = 0, \qquad \tilde{v} = -\frac{3+t^2}{54}K_0, \qquad \tilde{w} = \frac{12+t^2}{108}K_0, \tag{A.1}$$

is a particular solution of (4.5). Furthermore, u and v have same fundamental set as w (see (4.12)), and so by symmetry we have

$$\begin{aligned} u &= U^1\Psi_1 + U^3\Psi_3 + U^5\Psi^5 + U^7\Psi_7 + \tilde{u}, \\ v &= V^0\Psi_0 + V^2\Psi_2 + V^4\Psi_4 + V^6\Psi_6 + \tilde{v}, \\ w &= W^0\Psi_0 + W^2\Psi_2 + W^4\Psi_4 + W^6\Psi^6 + \tilde{w}. \end{aligned} \tag{A.2}$$

Applying the Euler equations (4.5a) and (4.5b), we can solve U^1, U^3, V^0, and V^2 as a function of W^0 and W^2 and, similarly, U^5, U^7, V^4, and V^6 as a function of W^4 and W^6. Finally W^0, W^2, W^4, and W^6 can be solved by claiming that (u, v, w) satisfies the boundary conditions (4.6a)–(4.6d). By symbolic calculus we obtain the expansions

$$\begin{aligned} W^0 &= \frac{1}{81}\left(\frac{3}{8}\right)^{1/4} K_0\sqrt{t} + \mathcal{O}(t), \\ V^0 &= -\frac{1}{2}W^0 + \mathcal{O}(t^2), \\ W^4 &= \frac{1}{18}\sqrt{\frac{3}{8}}K_0(\cos B_1 - \sin B_1)t + \mathcal{O}(t^{3/2}), \\ W^6 &= \frac{1}{18}\sqrt{\frac{3}{8}}K_0(\cos B_1 + \sin B_1)t + \mathcal{O}(t^{3/2}), \\ U^5 &= -\frac{1}{144}\left(\frac{8}{3}\right)^{1/4} K_0 \cos B_1\ t^{3/2} + \mathcal{O}(t^2), \\ U^7 &= \frac{1}{144}\left(\frac{8}{3}\right)^{1/4} K_0 \sin B_1\ t^{3/2} + \mathcal{O}(t^2), \end{aligned} \tag{A.3}$$

and the remaining coefficients are of order $\mathcal{O}(t^2)$.

Let $\underline{u} = \underline{u}^K - \underline{u}^0$. By (3.2) and (3.5), the α_2-independent part $\underline{u}_2^0$ of $\underline{u}^0$ is

$$\underline{u}_2^0 = \left(0, \quad -\frac{3}{54}K_0, \quad \frac{12}{108}K_0\right), \tag{A.4}$$

so by (A.1) through (A.4),

$$\mathcal{A}_\delta^K(\underline{u}, \underline{u}) = \mathcal{O}\big(t + \sqrt{t}\ e^{-\delta/\sqrt{t}}\big), \qquad \mathcal{B}_\delta^K(\underline{u}, \underline{u}) = \mathcal{O}\big(t^2 + \sqrt{t}\ e^{-\delta/\sqrt{t}}\big), \tag{A.5}$$

where $\mathcal{A}_\delta^K$ and $\mathcal{B}_\delta^K$ are defined in Remark 5.1 and the nonboundary layer term $\mathcal{O}(t)$ arises via $\kappa_{22}^K(\underline{u})$ because of the component $(u,v,w) = (0, V^0\Psi_0, W^0\Psi_0)$. Moreover, (A.5) implies that

$$|||\underline{u}|||_{K,t,\delta} = \mathcal{O}\big(\sqrt{t} + t^{1/4}e^{-\delta/\sqrt{t}}\big). \tag{A.6}$$

Finally, the leading term in the global relative error is

$$\frac{|||\underline{u}|||_{K,t}}{|||\underline{u}^K|||_{K,t}} = \frac{1}{3}\left(\frac{3}{8}\right)^{1/8} t^{1/4} + \mathcal{O}(\sqrt{t}) \approx 0.295\ t^{1/4}.$$

This is a result of the leading boundary layer

$$(u,v,w) = (U^5\Psi_5 + U^7\Psi_7,\ 0,\ W^4\Psi_4 + W^6\Psi_6).$$

(2) Let $f = K_1\alpha_1 \cos 2\alpha_2$. Then, by (3.2) and (3.5),

$$\underline{u}_2^0 = \left(-\frac{K_1}{72},\ -\frac{K_1\alpha_1}{36},\ \frac{K_1\alpha_1}{18}\right),$$

whereas a particular solution of (4.5) is

$$\tilde{u} = -\frac{12+7t^3}{432}K_1, \qquad \tilde{v} = -\frac{3+t^2}{54}K_1\alpha_1, \qquad \tilde{w} = \frac{12+t^2}{108}K_1\alpha_1.$$

The complete solution of (4.5) and (4.6) is then of the form

$$\begin{aligned} u &= U^0\Psi_0 + U^2\Psi_2 + U^4\Psi_4 + U^6\Psi_6 + \tilde{u},\\ v &= V^1\Psi_1 + V^3\Psi_3 + V^5\Psi_5 + V^7\Psi_7 + \tilde{v},\\ w &= W^1\Psi_1 + W^3\Psi_3 + W^5\Psi_5 + W^7\Psi_7 + \tilde{w}. \end{aligned}$$

Proceeding as above, we get

$$\begin{aligned} W^1 &= -\left(\frac{1}{18} - \frac{1}{108}\left(\frac{3}{8}\right)^{1/4}\sqrt{t}\right)K_1 + \mathcal{O}(t),\\ V^1 &= -\frac{1}{2}W^1 + \mathcal{O}(t^2),\\ U^0 &= \frac{1}{2}V^1 + \mathcal{O}(t^2),\\ W^5 &= \frac{1}{96}\sqrt{\frac{8}{3}}K_1(\cos B_1 + \sin B_1)t + \mathcal{O}(t^{3/2}),\\ W^7 &= \frac{1}{96}\sqrt{\frac{8}{3}}K_1(\cos B_1 - \sin B_1)t + \mathcal{O}(t^{3/2}),\\ U^4 &= \frac{1}{288}\left(\frac{8}{3}\right)^{1/4}K_1 \sin B_1\ t^{3/2} + \mathcal{O}(t^2),\\ U^6 &= -\frac{1}{288}\left(\frac{8}{3}\right)^{1/4}K_1 \cos B_1\ t^{3/2} + \mathcal{O}(t^2), \end{aligned}$$

and the remaining coefficients are again of order $\mathcal{O}(t^2)$. Also, this time it is easy to check that (A.5), and thus (A.6), hold, and now

$$\frac{|||\underline{u}|||_{K,t}}{|||\underline{u}^K|||_{K,t}} = \frac{1}{4}\left(\frac{8}{3}\right)^{3/8} t^{1/4} + \mathcal{O}(\sqrt{t}) \approx 0.361\, t^{1/4},$$

where the leading term comes from the leading boundary layer

$$(u, v, w) = (U^4\Psi_4 + U^6\Psi_6,\ 0,\ W^5\Psi_5 + W^7\Psi_7).$$

REFERENCES

[K] W. T. KOITER, *On the foundations of the linear theory of thin elastic shells*, Proc. Kon. Nederl. Akad. Wetensch., B 73 (1970), pp. 169–195.

[M] D. MORGENSTERN, *Herleitung der Plattentheorie aus der Dreidimensionalen Elastizitätstheorie*, Arch. Rational Mech. Anal., 4 (1959), pp. 145–152.

[NH] J. NEČAS AND I. HLAVACEK, *Mathematical Theory of Elastic and Elasto–Plastic Bodies, An Introduction*, Elsevier, Amsterdam, New York, 1981.

[N] V. V. NOVOZHILOV, *The Theory of Thin Shells*, P. G. Lowe, transl., J. R. M. Radok, ed., Noordhoff, Groningen, the Netherlands, 1959.

[P] J. PIILA, *Energy estimates and asymptotic analysis in the theory of shells*, Ph.D. thesis, preprint A320, 1993, Institute of Mathematics, Helsinki University of Technology.

[PP] J. PIILA AND J. PITKÄRANTA, *Energy estimates relating different elastic models of a thin cylindrical shell I. The membrane-dominated case*, SIAM J. Math. Anal., 24 (1993), pp. 1–22.

[P1] ———, *Energy estimates relating different linear elastic models of a cylindrical shell. (II) The bending-dominated case*, preprint A293, 1991, Institute of Mathematics, Helsinki University of Technology.

[P2] ———, *Energy estimates relating different linear elastic models of a cylindrical shell. (III) The soft membrane case*, preprint A299, 1991, Institute of Mathematics, Helsinki University of Technology.

[S] J. L. SANDERS, *An improved first-approximation theory of thin shells*, Tech. Report R–24, National Aeronautics and Space Administration, 1959.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 850–868, July 1995

004

A GLOBAL EXISTENCE AND UNIQUENESS THEOREM FOR A MODEL PROBLEM IN DYNAMIC ELASTO-PLASTICITY WITH ISOTROPIC STRAIN-HARDENING *

A. NOURI† AND M. RASCLE†

Abstract. We prove the global existence and uniqueness of the weak solution to the initial boundary value problem for an elastic-plastic solid with isotropic strain-hardening on a bounded domain in three space dimensions.

Key words. elastoplasticity, isotropic strain-hardening, evolution problem

AMS subject classification. 73E50

1. Introduction. In this paper, we consider a three-dimensional elastoplastic material, submitted to small strains motions. For such a material, we prove the existence and the uniqueness of a globally defined weak solution. In the case of an elastic perfectly plastic material, the quasistatic and dynamic evolution problems have been studied by several authors, see, e.g., [3], [11], [12]. For a material with isotropic hardening, we mention the paper of Laborde and Nguyen [9], in which they determine the evolution of the stress tensor, assuming the strain evolution is known. Here, we consider the full evolution problem. Roughly speaking, the associated operator has some monotone features, which of course correspond to its strongly dissipative properties, but is not monotone, at least not in a trivial way. This is due to the nonlinear function g, which appears in (2.4) below. Nevertheless, we take advantage of those monotone features. The paper is organized as follows. In §2, we provide some basic facts from the theory of plasticity and formulate the mathematical model to be studied. In §3, we give a precise statement of the problem, which naturally involves a convex set K that defines the plastic regime. In §4, we solve the regularized problem, essentially by studying the Yosida approximation of the subdifferential ∂I_K. The crucial point is to obtain the a priori estimate (4.20). In §5, we pass to the limit and solve the full problem. The key difficulty here is the estimate (5.38). We complete this paper in §6 with a short appendix on convex analysis.

2. Basic facts from the theory of plasticity. We start with Fig. 1, which describes a sequence of successive one-dimensional traction and compression tests for an elastoplastic material with isotropic strain-hardening. In this setting, as in the general case considered below, the body we consider is described in the reference configuration by a bounded domain Ω of $\mathbb{R}^n$ (here, $n = 1$ or 3) and is assumed to undergo small deformations; ε (resp., $\boldsymbol{\varepsilon}$) denotes the linearized strain (resp., strain tensor) and σ (resp., $\boldsymbol{\sigma}$) the stress (resp., stress tensor).

In Fig. 1, the material is initially in the elastic regime, until the stress σ reaches at point A the constant yield stress σ_Y. Then the regime becomes plastic until point B, where we start unloading elastically until point C. At point C, we start compressing elastically the body, until point D, where we reach the plastic compression regime.

* Received by the editors May 28, 1991; accepted for publication in revised form November 22, 1993.

†Département de Mathématiques, Université de Nice Sophia-Antipolis, Parc Valrose, 06108 Nice Cedex 2, France.

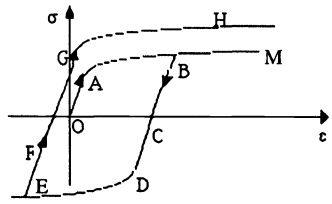


FIG. 1. *Isotropic hardening.*

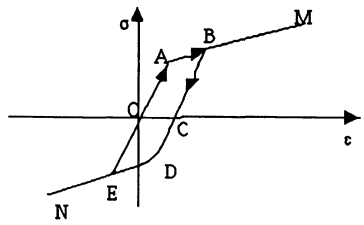


FIG. 2. *Kinematic hardening.*

Then we repeat the process, with a new elastic regime EFG, a new plastic traction regime GH, and so on. More precisely, we decompose the total strain ε into its elastic part ε_e and its plastic part π:

$$\varepsilon = \varepsilon_e + \pi \quad (\text{resp., } \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_e + \boldsymbol{\pi}) \tag{2.1}$$

The stress σ and the elastic strain ε_e are assumed to satisfy Hooke's law:

$$\varepsilon_e = \varepsilon - \pi = \frac{1}{E}\,\sigma, \tag{2.2}$$

$$\left(\text{resp., } \boldsymbol{\varepsilon}_e = \boldsymbol{\varepsilon} - \boldsymbol{\pi} = \frac{1+\nu}{E}\,\boldsymbol{\sigma} - \frac{\nu}{E}\,(\text{Tr } \boldsymbol{\sigma})\mathbf{1}\right). \tag{2.3}$$

Here, E and ν denote, respectively, the Young modulus and the Poisson coefficient; $(\text{Tr } \boldsymbol{\sigma}) = \sigma_{ii}$, the trace of the symmetric tensor $\boldsymbol{\sigma}$; and $\mathbf{1}$, the 3×3 identity tensor. In the elastic regime, the plastic strain π is locally constant. In the plastic regime, $\varepsilon, \varepsilon_e$, and π are given functions of the stress σ and of a hardening parameter β that describes the history of the material and specifies the yield curve (AB, DE, GH, and so on) that is actually involved. The first curve of importance in Fig. 1 is the phenomenological yield curve ABM, which we approximate by a given function g:

$$\sigma \overset{\text{def}}{=} \sigma_Y - g(\varepsilon_Y - \varepsilon). \tag{2.4}$$

A typical example of function g is

$$-g(\varepsilon_Y - \varepsilon) = k_Y(-(\varepsilon_Y - \varepsilon))^m, \qquad m \in (0,1). \tag{2.5}$$

The extreme cases $m = 0$ and $m = 1$ classically correspond, respectively, to a perfectly plastic material, where

$$|\sigma| \leq \sigma_Y, \tag{2.6}$$

and to the elastoplastic case, with kinematic strain-hardening. In the latter case (see Fig. 2), the family of successive yield curves reduces to a pair of straight lines ABM and DEN, which are symmetric with respect to the origin.

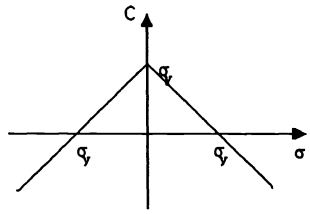


FIG. 3. *The convex set* $\bar{K}$.

For such a material, the size of the convex set of plasticity introduced below is constant, but its center evolves with respect to time. By contrast, for an elastoplastic material with isotropic strain-hardening, all these curves can be deduced from the "initial" curve ABM by a sequence of symmetries with respect to successive points, like points C, F, and so on. Thus, in Fig. 1, $|BC| = |CD|, |EF| = |FG|$, and so on. Therefore, for such a material, the size of the convex set of plasticity increases with respect to time, while its center remains on the horizontal axis $\sigma = 0$. One can combine these two types of models, see, e.g., [5]. We also remark that such models are rate-independent materials: the curves like ABCDEFGHI do not depend on the actual loading and unloading speeds. In contrast, the Yosida approximation, defined in §3, provides a family of mathematical rate-dependent approximations of those materials.

From now on, we consider an elastoplastic material with isotropic strain-hardening. We define the hardening parameter

$$\beta(x,t) = \int_0^t |\partial_t \pi|(x,s)\, ds \stackrel{\text{def}}{=} -\gamma(x,t), \tag{2.7}$$

where $|\ \ |$ denotes the absolute value. Observe that, by construction, γ is a decreasing nonpositive function of time.

In the one-dimensional case, we introduce the closed convex set

$$\bar{K} = \{(\sigma, C) \in \mathbb{R} \times \mathbb{R} / |\sigma| + C \leq \sigma_Y\}, \tag{2.8}$$

(see Fig. 3) and we define the convex set of plasticity K by

$$K = \{(\sigma, \gamma) \in \mathbb{R} \times \mathbb{R} / |\sigma| + g(\gamma) \leq \sigma_Y\}, \tag{2.9}$$

where $\beta = -\gamma$ is the hardening parameter defined in (2.7) and the function g defined in (2.4) describes the initial yield curve ABM.

Now, the elastic regime corresponds to

$$(\sigma, \gamma) \in \mathrm{Int}(K) : |\sigma| + g(\gamma) < \sigma_Y \tag{2.10}$$

or

$$(\sigma, \gamma) \in \partial K : |\sigma| + g(\gamma) = \sigma_Y \tag{2.11}$$

and

$$\partial_t |\sigma| \leq 0 \tag{2.12}$$

and

$$\partial_t \sigma = E\partial_t(\varepsilon - \pi) \tag{2.13}$$

and

$$\partial_t \pi = \partial_t \beta = 0, \tag{2.14}$$

while the plastic regime corresponds to

$$(\sigma, \gamma) \in \partial K : |\sigma| + g(\gamma) = \sigma_Y \tag{2.15}$$

and

$$\partial_t |\sigma| = (\operatorname{sgn} \sigma)\partial_t \sigma \geq 0 \tag{2.16}$$

and there exists $\partial \geq 0$ such that

$$\partial_t \gamma = -\partial_t \beta = -\lambda \tag{2.17}$$

$$\partial_t \pi = \lambda \operatorname{sgn}(\sigma). \tag{2.18}$$

Moreover, in either case,

$$\gamma(x,t) = -\beta(x,t) = \int_0^t |\partial_t \pi|(x,s)\, ds \leq 0. \tag{2.19}$$

Clearly, formulas (2.14), (2.15), (2.16), and (2.17) are contained in the general formula

$$\partial_t(\pi, -\gamma) \in \partial I_{\bar{K}}((\sigma, g(\gamma)), \tag{2.20}$$

where $I_{\bar{K}}$ is the indicator function of the closed convex set $\bar{K}$ and $\partial I_{\bar{K}}$ its subdifferential, see, e.g., [4]. These classical notions are briefly recalled in §6. We just note here that

$$\partial I_{\bar{K}}((\sigma, g(\gamma))) = \begin{cases} \{0\} & \text{if } (\sigma,\gamma) \in \operatorname{Int}(K), \\ \varnothing & \text{if } (\sigma,\gamma) \notin K, \\ \{\lambda(\operatorname{sgn}\sigma, 1); \lambda \geq 0\} & \text{if } (\sigma,\gamma) \in \partial K \quad \text{and} \quad g(\gamma) \leq 0. \end{cases} \tag{2.21}$$

We have defined K as the convex set of plasticity. Indeed, K is convex if the function g is convex, i.e., by (2.4), if the curve ABM in Fig. 1 is convex downward, which we assume from now on.

3. Statement of the problem.

3.1. The one-dimensional version. We first consider a more tractable one-dimensional version, in the spirit of §2. The yield function g appearing in (2.4) is assumed to satisfy the following assumptions:

(H1) g is a *convex,* increasing, smooth function from $]-\infty, 0]$ into $]-\infty, 0]$, such that $g(0) = 0$. Then g is extended in a C^1 fashion to $[0, +\infty[$ by a linear relation.

(H2) $\exists \alpha > 0, \exists \beta > 0 / \forall \gamma \in \mathbb{R}, 0 < \alpha \leq g'(\gamma) \leq \beta < E$.

Remark 3.1. Assumption (H2), which implies a linear growth of the function g at infinity, is in fact quite natural, since g is *convex.* As we already said, the set of

plasticity is indeed convex if (and only if) g is convex. Moreover, it is worthwhile to note that our results are no longer true if g is concave, since in this case there are solutions that definitely contain shock waves in the plastic regime, see e.g., Antman [1], Trangenstein and Pember [15]. However, the solutions constructed in the abovementioned papers are essentially solutions to the Riemann problem and therefore already contain discontinuities at time $t = 0$; it is likely, but not entirely clear, that shock waves would develop in finite time, even starting with smooth initial data. In contrast, with a convex yield function g as here, the solution of the Riemann problem only involves contact discontinuities (in the elastic regime) and rarefaction waves (in the plastic regime). Therefore, under assumptions (H1) and (H2), on one hand the *propagation* of plastic shock waves is impossible, and on the other hand, as we are going to show in the next sections, solutions in Sobolev spaces are globally defined, which implies that no plastic shock wave can *develop* if we start with smooth initial data.

Our material is now a one-dimensional elastoplastic bar, with isotropic strain-hardening, defined in the reference configuration by a bounded interval $\Omega =]0, L[$, undergoing small deformations. We denote by v the velocity, the density ρ is supposed to be constant, and we assume that there is no external load (see a comment on this assumption in Remark 3.2 below). For simplicity, we also assume that $\rho = 1$ and $E = 1$.

We now write the equations. First, the fundamental law of mechanics gives

$$\partial_t v - \partial_x \sigma = 0. \tag{3.1}$$

The compatibility of second order derivatives of the displacement implies

$$\partial_t \varepsilon - \partial_x v = 0. \tag{3.2}$$

It is more convenient to introduce the vector-valued functions

$$U = (v, \sigma, \gamma) \in \mathbb{R}^3, \tag{3.3}$$

$$\bar{U} \stackrel{\text{def}}{=} G(U) = (v, \sigma, C) \stackrel{\text{def}}{=} (v, \sigma, g(\gamma)) \in \mathbb{R}^3 \tag{3.4}$$

to slightly modify the convex set of plasticity

$$K = \{U = (v, \sigma, \gamma) \in \mathbb{R}^3 / |\sigma| + g(\gamma) \leq 0\} \tag{3.5}$$

and to introduce the sets

$$K_- = \{U \in K / \gamma \leq 0\}, \tag{3.6}$$

$$\bar{K} = \{\bar{U} = (v, \sigma, C) \in \mathbb{R}^3 / |\sigma| + C \leq 0\} = G(K), \tag{3.7}$$

$$\bar{K}_- = \{\bar{U} \in \bar{K} / C \leq 0\} = G(K_-). \tag{3.8}$$

Since those definitions do not involve the velocity v, the corresponding sets are cylinders, parallel to the v-axis. Therefore the corresponding $\partial I_{\bar{K}}$ have no v-component.

We can now eliminate $\varepsilon = \sigma + \pi$ (here, $E = 1$) and rewrite the system in the form

$$0 \in (\partial_t v - \partial_x \sigma, \partial_t \sigma - \partial_x v, \partial_t \gamma) + \partial I_{\bar{K}}(v, \sigma, g(\gamma)), \tag{3.9}$$

i.e.,

$$0 \in \partial_t AU + BU + \partial I_{\bar{K}}(G(U)), \tag{3.10}$$

where

$$AU \stackrel{\text{def}}{=} U; \qquad BU \stackrel{\text{def}}{=} (-\partial_x \sigma, -\partial_x v, 0). \tag{3.11}$$

The inequality $\gamma \leq 0$ will be satisfied a posteriori.

We add classical boundary conditions, e.g.,

$$\sigma(0,t) \equiv 0, \qquad v(L,t) \equiv 0. \tag{3.12}$$

Of course, (3.12) implies

$$(v \cdot \sigma)(0,t) \equiv (v \cdot \sigma)(L,t) \equiv 0. \tag{3.13}$$

Finally, we add the initial data

$$(v, \sigma, \gamma)(x,0) = (v_0(x), \sigma_0(x), 0) \in K \quad \text{a.e. in } \Omega =]0, L[. \tag{3.14}$$

Therefore, in one space dimension, our initial boundary value problem (IBVP_1) is precisely stated in formulas (3.10) and (3.12)–(3.14) above.

3.2. The full three-dimensional problem. The function g still satisfies assumptions (H1) and (H2). The modifications are the following:

- $\boldsymbol{\sigma} \in M_s$, the set of symmetric 3×3 tensors, equipped with its classical scalar product $\boldsymbol{\sigma}:\boldsymbol{\tau}$ and the associated norm $|\boldsymbol{\sigma}|$.
- $\boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon}_e = \boldsymbol{\varepsilon} - \boldsymbol{\pi}$ are related by Hooke's law (2.2).
- $\partial_t \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(v) = (\frac{1}{2}(\partial_i v_j + \partial_j v_i))_{1 \leq i,j \leq 3}$, where $\partial_i \stackrel{\text{def}}{=} \partial/\partial x_i$.
- we classically decompose the total stress $\boldsymbol{\sigma}$ into the (trace-free) stress deviator $\boldsymbol{\sigma}^D$ and the hydrostatic pressure tensor

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^D - p\mathbf{1}. \tag{3.15}$$

The convex set of plasticity is now

$$K = \{U = (v, \boldsymbol{\sigma}, \gamma) \in \mathbb{R}^3 \times M_s \times \mathbb{R} / |\boldsymbol{\sigma}^D| - \sigma_y + g(\gamma) \leq 0\}, \tag{3.16}$$

and we define $\bar{U}$ and $\bar{K}$,

$$\bar{K} = \{\bar{U} = (v, \boldsymbol{\tau}, C) \in \mathbb{R}^3 \times M_s \times \mathbb{R} / |\boldsymbol{\tau}| - \sigma_y + C \leq 0\}, \tag{3.17}$$

so that

$$K = \{U = (v, \boldsymbol{\sigma}, \gamma) \in \mathbb{R}^3 \times M_s \times \mathbb{R} / \bar{U} = G(U) \in \bar{K}\}, \tag{3.18}$$

with

$$G(U) = (v, \boldsymbol{\sigma}^D, g(\gamma)) = \left(v, \boldsymbol{\sigma} - \frac{1}{3}(\mathrm{Tr}\boldsymbol{\sigma})\mathbf{1}, g(\gamma)\right). \tag{3.19}$$

The system of equations is now

$$0 \in \left(\partial_t v - \mathrm{div}\boldsymbol{\sigma}, \frac{1+\nu}{E}\,\partial_t\boldsymbol{\sigma} - \frac{\nu}{E}\,\mathrm{Tr}(\partial_t\boldsymbol{\sigma})\mathbf{1} - \boldsymbol{\varepsilon}(v), \partial_t\gamma\right) + \partial I_{\bar{K}}(v, \boldsymbol{\sigma}, g(\gamma)). \tag{3.20}$$

Let N denote the fourth-order tensor of the linear elasticity system, such that $\boldsymbol{\varepsilon}^e = N\boldsymbol{\sigma}$, (see (2.3)). Let us define

$$\begin{cases} V \overset{\text{def}}{=} AU \overset{\text{def}}{=} (v, \boldsymbol{\varepsilon}^e, \gamma) = (v, N\boldsymbol{\sigma}, \gamma), \\ BU \overset{\text{def}}{=} (-\mathrm{div}\boldsymbol{\sigma}, -\boldsymbol{\varepsilon}(v), 0), \\ CU \overset{\text{def}}{=} \partial I_{\bar{K}}(G(U)). \end{cases} \tag{3.21}$$

We can rewrite (3.20) in the form

$$0 \in \partial_t AU + BU + CU, U = (v, \boldsymbol{\sigma}, \gamma)$$

and look for a solution in the Hilbert space,

$$H = L^2(\Omega)^3 \times M_s(L^2(\Omega)) \times L^2(\Omega), \tag{3.22}$$

where $M_s(L^2(\Omega))$ is the space of symmetric 3×3 tensor-valued functions. We equip H with the scalar-product

$$(U, U^*) \overset{\text{def}}{=} ((v, \boldsymbol{\sigma}, \gamma), (v^*, \boldsymbol{\sigma}^*, \gamma^*)), = \int_\Omega (v \cdot v^* + \boldsymbol{\sigma} : \boldsymbol{\sigma}^* + \gamma\gamma^*)(x)\, dx. \tag{3.23}$$

We add natural boundary conditions and initial data. Finally, the problem we consider is

$$(\mathcal{P}) \begin{cases} \text{find } U \in L^\infty(0,T;H) \text{ such that} & \\ 0 \in \partial_t AU + BU + CU, \qquad U = (v, \boldsymbol{\sigma}, \gamma), & (3.24) \\ \boldsymbol{\sigma} \cdot n \equiv 0 \quad \text{on } \partial\Omega_1, \qquad v \equiv 0 \quad \text{on } \partial\Omega_2, \quad \forall t > 0, & (3.25) \\ (v, \boldsymbol{\sigma}, \gamma)(x, 0) \equiv (v_0(x), \boldsymbol{\sigma}_0(x), 0) \in K \quad \text{a.e. in } \Omega, & (3.26) \end{cases}$$

where A, B, C are defined in (3.21) and $(\partial\Omega_1, \partial\Omega_2)$ is a partition of the (piecewise) smooth boundary $\partial\Omega$.

Remark 3.2. (i) We could naturally consider inhomogeneous boundary conditions, at least if they remain "elastic." The plastic case would require a more careful analysis and the regularity assumptions that are specified below could be violated.

(ii) Similarly, it would be more realistic to consider a general (given) nonzero external load $f(x,t)$, e.g., such that f and $\partial f/\partial t$ lie in the space $L^\infty(0,T;L^2(\Omega)^3)$. In this case, the right-hand side of relations (4.5a) and (4.8a) below would be replaced, respectively, by f and $\frac{\partial f}{\partial t}$, which would modify the energy estimate (4.12) (see also (5.4)) in an obvious way and would still provide the same type of estimates by Gronwall's lemma (observe that only the first equation—the easiest one—would be modified in a system like (4.8)). So in principle our results should also apply to this case.

(iii) Again the inequality $\gamma \leq 0$ will be satisfied a posteriori.

(iv) It would be equivalent to invert the relationship

$$\boldsymbol{\varepsilon}^e = N\boldsymbol{\sigma}$$

to define

$$U \overset{\text{def}}{=} A^{-1}V = (v, N^{-1}\boldsymbol{\varepsilon}^e, \gamma) = (v, \boldsymbol{\sigma}, \gamma)$$

and to rewrite the problem under the form

$$(\mathcal{P}') \begin{cases} \text{Find } V \in L^\infty(0,T;H) \text{ such that} & \\ 0 \in \partial_t V + B(A^{-1}V) + C(A^{-1}V); \qquad V = (v, \boldsymbol{\varepsilon}^e, \gamma), & (3.27) \\ N^{-1}\boldsymbol{\varepsilon}^e \cdot n \equiv 0 \quad \text{on } \partial\Omega_1; \qquad v \equiv 0 \quad \text{on } \partial\Omega_2 \forall t > 0, & (3.28) \\ (v, \boldsymbol{\varepsilon}^e, \gamma)(x,0) \equiv (v_0(x), \boldsymbol{\varepsilon}^e_0(x), 0) \quad \text{on } \Omega. & (3.29) \end{cases}$$

But then we would have to change the scalar product in H:

$$\begin{aligned} [V, V^*] &= [(v, \boldsymbol{\varepsilon}^e, \gamma), (v^*, \boldsymbol{\varepsilon}^{e*}, \gamma^*)] \\ &= \int_\Omega (v \cdot v^* + (N^{-1}\boldsymbol{\varepsilon}^e) : (N^{-1}\boldsymbol{\varepsilon}^{e*}) + \gamma\gamma^*)(x)\, dx \\ &= \int_\Omega (v \cdot v^* + \boldsymbol{\sigma} : \boldsymbol{\sigma}^* + \gamma\gamma^*)(x)\, dx \\ &= (A^{-1}V, A^{-1}V^*). \end{aligned} \tag{3.30}$$

From now on, we consider either the problem $(\mathcal{P})$ or the equivalent problem $(\mathcal{P}')$.

4. Regularization of the three-dimensional problem. As we have seen, A or A^{-1} can be easily inverted in a positive definite way from H into H. As we will see below, B is a maximal monotone operator on the Hilbert space H. However, due to the nonlinearity of function g, the operator

$$U \to CU = \partial I_{\bar{K}}(G(U))$$

is not monotone, even if we try to change the scalar product. Therefore, in a first step we are going to replace $\partial I_{\bar{K}}$ by its Yosida approximation, see, e.g., [4],

$$(\partial I_{\bar{K}})_\mu(\bar{U}) = \frac{1}{\mu}(I - P_{\bar{K}})(\bar{U}). \tag{4.1}$$

Therefore, we replace CU by

$$C_\mu U = \frac{1}{\mu}(I - P_{\bar{K}})(G(U)), \tag{4.2}$$

where μ decreases to 0_+, I is the identity, and $P_{\bar{K}}$ is defined pointwise as the orthogonal projection on the convex set

$$\bar{K} = \{\bar{U} = (v, \boldsymbol{\tau}, C) \in \mathbb{R}^3 \times M_s \times \mathbb{R} / |\boldsymbol{\tau}| - \sigma_y + C \leq 0\}.$$

Using general results on convex sets defined pointwise, see, e.g., [7], it is easy to check the following lemma.

LEMMA 4.1. (i) *For all* $U = (v, \boldsymbol{\sigma}, \gamma)$ *in* $\mathbb{R}^3 \times M_s \times \mathbb{R}$,

$$C_\mu U = \begin{cases} 0 & \text{if } |\boldsymbol{\sigma}^D| - \sigma_y + g(\gamma) \le 0, \\ \dfrac{1}{2\mu}\,(|\boldsymbol{\sigma}^D| - \sigma_y + g(\gamma))\left(0, \dfrac{\boldsymbol{\sigma}^D}{|\boldsymbol{\sigma}^D|}, 1\right) & \text{if } |\boldsymbol{\sigma}^D| - \sigma_y + g(\gamma) > 0, \end{cases} \tag{4.3}$$

$$= \frac{1}{2\mu}\,(|\boldsymbol{\sigma}^D| - \sigma_y + \gamma(\gamma))_+ \left(0, \frac{\boldsymbol{\sigma}^D}{|\boldsymbol{\sigma}^D|}, 1\right).$$

(ii) *The same formula defines* $C_\mu U$ *for any* $U = (v, \boldsymbol{\sigma}, \gamma)$ *in* H. *Here,* $\xi_+ \overset{\text{def}}{=} \max(\xi, 0)$ *is the positive part of any real number* ξ.

4.1. Use of the Yosida approximation of ∂I_K. For each fixed positive μ, C_μ is a Lipschitz operator, and the full operator is a Lipschitz perturbation of a maximal monotone operator.

THEOREM 4.1. (i) *Assume* $(v_0, \boldsymbol{\sigma}_0, 0) \in D(B)$, *the domain of* B, *with* $|\boldsymbol{\sigma}_0^D| \le \sigma_y$ *almost everywhere in* Ω. *Then, for each* $\mu > 0$, *the regularized problem*

$$\partial_t AU + BU + \frac{1}{\mu}\,(I - P_{\bar{K}})(U) = 0, \tag{4.4}$$

i.e.,

$$(\mathcal{P}_\mu) \begin{cases} \partial_t v - \operatorname{div}\boldsymbol{\sigma} = 0, & \text{(4.5a)} \\ \partial_t \boldsymbol{\varepsilon}^e - \boldsymbol{\varepsilon}(v) = -\dfrac{1}{2\mu}\,(|\boldsymbol{\sigma}^D| - \sigma_y + g(\gamma))_+ \dfrac{\boldsymbol{\sigma}^D}{|\boldsymbol{\sigma}^D|}, & \text{(4.5b)} \\ \partial_t \gamma = -\dfrac{1}{2\mu}\,(|\boldsymbol{\sigma}^D| - \sigma_y + g(\gamma))_+ \le 0, & \text{(4.5c)} \end{cases}$$

together with the boundary conditions (3.25) *and initial data* (3.26), *has a unique solution* $(v, \boldsymbol{\sigma}, \gamma) \overset{\text{def}}{=} (v^\mu, \boldsymbol{\sigma}^\mu, \gamma^\mu)$ *in the space* $W^{1,\infty}(0, T; H)$.

(ii) *Moreover,* $\gamma^\mu \le 0$ *and* $U^\mu = (v^\mu, \boldsymbol{\sigma}^\mu, \gamma^\mu)$ *is bounded in this space, uniformly with respect to* μ. *Similarly,* $\boldsymbol{\varepsilon}(v^\mu)$ *and* $\operatorname{div}\boldsymbol{\sigma}^\mu$ *are uniformly bounded in* $L^\infty(0, T; M_s(L^2(\Omega))$ *and* $L^\infty(0, T; L^2(\Omega))$, *respectively. Finally, the right-hand side in* (4.5b), (4.5c) *is also bounded in the same spaces, respectively.*

Proof of Theorem 4.1. (i) We first note that B is a skew symmetric operator on H, with domain

$$D(B) = \{U = (v, \boldsymbol{\sigma}, \gamma) \in H / \varepsilon(v) \in M_s(L^2(\Omega)), \\ v/_{\partial\Omega_2} = 0, \operatorname{div}(\boldsymbol{\sigma}) \in (L^2(\Omega))^3, \boldsymbol{\sigma} \cdot n/_{\partial\Omega_1} = 0\}. \tag{4.6}$$

Using the Korn inequality,

$$D(B) = \{U = (v, \boldsymbol{\sigma}, \gamma) \in (H^1(\Omega))^3 \times M_s(L^2(\Omega)) \times L^2(\Omega) / v/_{\partial\Omega_2} = 0, \\ \operatorname{div}(\boldsymbol{\sigma}) \in (L^2(\Omega))^3, \boldsymbol{\sigma} \cdot n/_{\partial\Omega_1} = 0). \tag{4.7}$$

Therefore $B = -B^*$ is monotone, and its domain is dense in H. Since

$$\forall U \in D(B), |U| \le |(I + B)^* U|,$$

we deduce that $I + B$ is surjective. Therefore, the operator B is maximal monotone on H. Of course, the same result is true for the operator $\{BA^{-1} : V \to B(A^{-1}V) = BU\}$, if we replace the scalar product in H by $[\cdot, \cdot]$, as defined in (3.30). Therefore, for each

$\mu > 0$, $(\mathcal{P}_\mu)$ is the Cauchy problem for a Lipschitz perturbation of a maximal monotone operator BA^{-1}, with initial data in the domain $D(BA^{-1})$.

By a classical result, see, e.g., [4], there exists a unique solution to $(\mathcal{P}_\mu)$, such that

$$V^\mu = (v^\mu, \varepsilon^{e\mu}, \gamma^\mu) = (v^\mu, N\sigma^\mu, \gamma^\mu) \in W^{1,\infty}(0,T;H).$$

Naturally, when μ decreases to 0_+, the norm of V^μ (or U^μ) in the space $W^{1,\infty}(0,T;H)$ could blow up as $e^{LT/\mu}$, where L is the Lipschitz constant of $(I - P_{\bar{K}})$ (here, $L = 1$).

(ii) This is not the case. Let us show part (ii) of Theorem 4.1. For each fixed $\mu > 0$, let us derive (4.5) with respect to t:

$$\begin{cases} \partial_t(\partial_t v^\mu) - \operatorname{div}(\partial_t \sigma^\mu) = 0, & \text{(4.8a)} \\ \partial_t(N\partial_t\sigma^\mu) - \varepsilon(\partial_t v^\mu) \overset{\text{def}}{=} \partial_t\left(h^\mu \dfrac{\sigma^{D\mu}}{|\sigma^{D\mu}|}\right) \overset{\text{def}}{=} k^\mu, & \text{(4.8b)} \\ \partial_t(\partial_t\gamma^\mu) \overset{\text{def}}{=} \partial_t(h^\mu), & \text{(4.8c)} \end{cases}$$

where

$$h^\mu \overset{\text{def}}{=} -\frac{1}{2\mu}\,(|\sigma^{D\mu}| - \sigma_y + g(\gamma^\mu))_+. \tag{4.9}$$

Since $U^\mu \in W^{1,\infty}(0,T;H)$, (4.8) is a linear system, where the unknown functions $\partial_t v^\mu, \partial_t\sigma^\mu, \partial_t\gamma^\mu$, and the right-hand side are in $L^\infty(0,T;H)$. Note that (4.8a) and (4.8b) are the system of linear elasticity. Moreover, the boundary conditions and initial data satisfy

$$\partial_t\sigma^\mu \cdot n \equiv 0 \text{ on } \partial\Omega_1, \qquad \partial_t v^\mu \equiv 0 \quad \text{on } \partial\Omega_2, \quad \forall t > 0, \tag{4.10}$$

$$\begin{cases} \partial_t v^\mu(x,0) \equiv \operatorname{div}(\sigma^\mu)(x,0) \in (L^2(\Omega))^3, \\ N\partial_t\sigma^\mu(x,0) \equiv \varepsilon(v^\mu)(x,0) \in M_s(L^2(\Omega)), \\ \partial_t\gamma^\mu(x,0) \in L^2(\Omega). \end{cases} \tag{4.11}$$

Let us consider (4.8a) and (4.8b). We know from the theory of linear elasticity that the following energy estimate is satisfied:

$$\begin{aligned} &\frac{1}{2}\left[\int_\Omega \partial_t v^\mu \cdot \partial_t v^\mu(x,s)\,dx\right]_{s=0}^{s=t} + \frac{1}{2}\left[\int_\Omega \partial_t\sigma^\mu : (N\partial_t\sigma^\mu)\,dx\right]_{s=0}^{s=t} \\ &\qquad = \int_0^t \int_\Omega k^\mu : \partial_t\sigma^\mu(x,s)\,dx\,ds. \end{aligned} \tag{4.12}$$

Here, we have respectively multiplied (4.8a) and (4.8b) by $\partial_t v^\mu$ and $\partial_t\sigma^\mu$ and used the boundary conditions, which imply

$$\int_0^t \int_\Omega (\operatorname{div}(\partial_t\sigma^\mu) \cdot \partial_t v^\mu + \partial_t\sigma^\mu : \varepsilon(\partial_t v^\mu))(x,s)\,dx\,ds = \int_0^t \int_{\partial\Omega} (\sigma^\mu \cdot n) \cdot v^\mu\,dS\,ds = 0. \tag{4.13}$$

On the other hand, the right-hand side of (4.8) is the time derivative of a Lipschitz function of U^μ, where $U^\mu \in W^{1,\infty}(0,T;H)$. Therefore, dropping the index μ and using the chain-rule formula in (4.9), we have

$$\partial_t\left(h\frac{\sigma^D}{|\sigma^D|}\right) = \partial_t h\left(\frac{\sigma^D}{|\sigma^D|}\right) - \left(\frac{h}{|\sigma^D|^3}\right)(\sigma^D : \partial_t\sigma^D)\sigma^D + \left(\frac{h}{|\sigma^D|}\right)\partial_t\sigma^D, \tag{4.14}$$

where

$$\partial_t h = -(2\mu)^{-1}\mathrm{sgn}_+(|\boldsymbol{\sigma}^D| + g(\gamma) - \sigma_y)\left(\frac{\boldsymbol{\sigma}^D}{|\boldsymbol{\sigma}^D|} : \partial_t\boldsymbol{\sigma}^D + g'(\gamma)\partial_t\gamma\right) \tag{4.15}$$

and

$$\mathrm{sgn}_+(\xi) \overset{\text{def}}{=} (\mathrm{sgn}\xi)_+.$$

Now, let us take the scalar product of (4.14) with $\partial_t\boldsymbol{\sigma}$, multiply (4.15) by $g'(\gamma)\partial_t\gamma$, and add. Since, for instance,

$$\begin{aligned}\partial_t\boldsymbol{\sigma}^D : \boldsymbol{\sigma} &= \partial_t\boldsymbol{\sigma}^D : \left(\boldsymbol{\sigma}^D + \frac{1}{3}\mathrm{Tr}(\boldsymbol{\sigma})\mathbf{1}\right)\\ &= \partial_t\boldsymbol{\sigma}^D : \boldsymbol{\sigma}^D + \frac{1}{3}\mathrm{Tr}(\boldsymbol{\sigma})(\partial_t\boldsymbol{\sigma}^D : \mathbf{1})\\ &= \partial_t\boldsymbol{\sigma}^D : \boldsymbol{\sigma}^D,\end{aligned} \tag{4.16}$$

we have

$$\begin{aligned}R &\overset{\text{def}}{=} \partial_t\boldsymbol{\sigma} : \partial_t\left(h\frac{\boldsymbol{\sigma}^D}{|\boldsymbol{\sigma}^D|}\right) + g'(\gamma)\partial_t\gamma\partial_t h\\ &= \partial_t h\left(\tau_t\boldsymbol{\sigma}^D : \left(\frac{\boldsymbol{\sigma}^D}{|\boldsymbol{\sigma}^D|}\right) + g'(\gamma)\partial_t\gamma\right) - \left(\frac{h}{|\boldsymbol{\sigma}^D|^3}\right)\\ &\quad\times((\boldsymbol{\sigma}^D : \partial_t\boldsymbol{\sigma}^D)^2 - |\boldsymbol{\sigma}^D|^2|\partial_t\boldsymbol{\sigma}^D|^2).\end{aligned} \tag{4.17}$$

By the Cauchy–Schwarz inequality, the second term in (4.17) is nonpositive. On the other hand, using (4.15),

$$R \le -(2\mu)^{-1}\mathrm{sgn}_+(|\boldsymbol{\sigma}^D| + g(\gamma) - \sigma_y)\left(\frac{\boldsymbol{\sigma}^D}{|\boldsymbol{\sigma}^D|} : \partial_t\boldsymbol{\sigma}^D + g'(\gamma)\partial_t\gamma\right)^2 \le 0. \tag{4.18}$$

Now, let us multiply (4.8c) by $g'(\gamma)\partial_t\gamma$, integrate on $\Omega\times[0,t]$, and add to (4.12). We obtain

$$\begin{aligned}&\frac{1}{2}\left[\int_\Omega |\partial_t|v^\mu|^2 + \partial_t\boldsymbol{\sigma}^\mu : (N\partial_t\boldsymbol{\sigma}^\mu)(x,s)\,dx\right]_{s=0}^{s=t} + \int_0^t\int_\Omega g'(\gamma)\partial_t\gamma\partial_t^2\gamma\,dx\,ds\\ &\quad= \int_0^t\int_\Omega R\,dx\,ds \le 0.\end{aligned} \tag{4.19}$$

On the other hand, since $g'' \ge 0$ and $\partial_t\gamma \le 0$,

$$\begin{aligned}\int_0^t\int_\Omega g'(\gamma)\partial_t\gamma\partial_t^2\gamma\,dx\,ds &= \left[\int_\Omega g'(\gamma)\frac{1}{2}(\partial_t\gamma)^2(x,s)\,dx\right]_{s=0}^{s=t}\\ &\quad - \int_0^t\int_\Omega g''(\gamma)\frac{1}{2}(\partial_t\gamma)^3(x,s)\,dx\,ds\\ &\ge \left[\int_\Omega g'(\gamma)\frac{1}{2}(\partial_t\gamma)^2(x,s)\,dx\right]_{s=0}^{s=t}\\ &\ge \frac{\alpha}{2}\left[\int_\Omega (\partial_t\gamma)^2(x,s)\,dx\right]_{s=0}^{s=t},\end{aligned} \tag{4.20}$$

where α appears in assumption (H2).

Clearly, the convexity of the yield function g is essential to obtain this estimate (4.20), which was indeed the crucial point. Combining (4.20) with the other important estimate (4.19), we obtain

$$\begin{aligned}
&\int_\Omega \left(\frac{1}{2}|\partial_t v^\mu|^2 + \partial_t \boldsymbol{\sigma}^\mu : (N\partial_t \boldsymbol{\sigma}^\mu) + \frac{1}{2}g'(\gamma^\mu)(\partial_t \gamma^\mu)^2\right)(x,t)\,dx \\
&\qquad \le \int_\Omega \left(\frac{1}{2}|\partial_t v^\mu|^2 + \partial_t \boldsymbol{\sigma}^\mu : (N\partial_t \boldsymbol{\sigma}^\mu) + \frac{1}{2}g'(0)(\partial_t \gamma^\mu)^2\right)(x,0)\,dx \\
&\qquad = \int_\Omega \left[\frac{1}{2}|\mathrm{div}\boldsymbol{\sigma}_0|^2 + (\boldsymbol{\varepsilon}(v_0) : N^{-1}\boldsymbol{\varepsilon}(v_0))\right](x)\,dx.
\end{aligned} \tag{4.21}$$

Therefore, the sequence (U^μ) is uniformly bounded in $W^{1,\infty}(0,T;H)$.

Finally, since

$$|\partial_t(N\boldsymbol{\sigma}^\mu) - \boldsymbol{\varepsilon}(v^\mu)| = |\partial_t \gamma^\mu|$$

and

$$\mathrm{div}(\boldsymbol{\sigma}^\mu) = 2_t v^\mu,$$

the uniform boundedness of $\partial_t U^\mu$ in $L^\infty(0,T;H)$ implies that $\boldsymbol{\varepsilon}(v^\mu)$ and $\mathrm{div}(\boldsymbol{\sigma}^\mu)$ are also uniformly bounded in $L^\infty(0,T;M_s(L^2(\Omega))$ (resp., $L^\infty(0,T;(L^2(\Omega))^3)$). For the same reason, the right-hand side in (4.5b) (resp., (4.5c)) is also uniformly bounded in the same space (resp., in $L^\infty(0,T;L^2(\Omega))$). Therefore, when $\mu \to 0$,

$$\|(I - P_{\bar{K}})(\bar{U}^\mu)\|_{L^\infty(0,T;H)} = O(\mu). \tag{4.22}$$

5. The main result.

THEOREM 5.1. *The sequence* $(v^\mu, \boldsymbol{\sigma}^\mu, \gamma^\mu)$ *defined in Theorem* 4.1 *satisfies*
(i)

$$v^\mu \to v \text{ in } L^p(0,T;(L^2(\Omega))^3) \text{ strongly, } \forall p < +\infty,$$

$$\boldsymbol{\sigma}^\mu \to \boldsymbol{\sigma} \text{ in } L^p(0,T;M_s(L^2(\Omega))) \text{ strongly, } \forall p < +\infty,$$

$$\gamma^\mu \to \gamma \text{ in } L^p(0,T;L^2(\Omega)) \text{ strongly, } \forall p < +\infty;$$

(ii) *the limit* $U = (v, \boldsymbol{\sigma}, \gamma)$, *globally defined, is the unique weak solution to the problem* $(\mathcal{P})$, *in the following sense*:

$$\begin{aligned}
&\forall \bar{V} \in L^1(0,T;H)/\bar{V}(x,t) \in \bar{K} \quad \text{a.e. in } \Omega x(0,T), \\
&\int_0^T ((\partial_t(AU) + BU)(t), (G(U) - \bar{V})(t))_H\,dt \le 0.
\end{aligned} \tag{5.1}$$

In particular, $U(x,t)$ *lies almost everywhere in* K *and* $\gamma \le 0$ *almost everywhere. Moreover,*

$$v \in L^\infty(0,T;(H^1(\Omega))^3) \cap W^{1,\infty}(0,T;L^2(\Omega)),$$

$$\mathrm{div}\boldsymbol{\sigma} \in L^\infty(0,T;(L^2(\Omega))^3),$$

$$\partial_t \boldsymbol{\sigma} \in L^\infty(0,T;M_s(L^2(\Omega))),$$

$$\partial_t \gamma \in L^\infty(0,T;L^2(\Omega)).$$

Remark 5.1. In (5.1), the upper bound T in the integral can be easily replaced by almost all t in $(0,T)$. It would be equivalent to ask that, almost everywhere, in $(0,T)$, for any $\bar{V} = \bar{V}(x) \in K$,

$$((\partial_t(AU) + BU)(t), (G(U) - \bar{V})(t))_H \leq 0. \tag{5.1'}$$

Proof. The proof consists of four steps. In the first step, we define the weak-star limits of the various subsequences. In the second step, we derive the convexity inequalities between these limits and define the boundary conditions and initial data satisfied by the limit. In Step 3, we prove that the weak-star limit is a solution. Finally, in Step 4, we prove that the convergence is strong and the solution is unique. Therefore, the full sequence converges strongly to the unique solution.

Step 1. By definition of the orthogonal projection on a convex set, for each $\mu > 0, \bar{U}^\mu = G(U^\mu)$ satisfies almost everywhere

$$\forall \bar{V} \in \bar{K}, \quad (\bar{U}^\mu - P_{\bar{K}}\bar{U}^\mu, P_{\bar{K}}\bar{U}^\mu - \bar{V}) \geq 0.$$

From (4.4), this implies

$$\text{(5.2a)} \quad \forall \bar{V} \in \bar{K}, \quad (\partial_t AU^\mu + BU^\mu, P_{\bar{K}}\bar{U}^\mu - \bar{V}) = -\mu^{-1}(\bar{U}^\mu - P_{\bar{K}}\bar{U}^\mu, P_{\bar{K}}\bar{U}^\mu - \bar{V}) \leq 0.$$

On the other hand, (4.4) also implies

$$(\partial_t AU^\mu + BU^\mu, G(U^\mu) - P_{\bar{K}}(G(U^\mu))) = -\mu^{-1}\|\bar{U}^\mu - P_{\bar{K}}\bar{U}^\mu\|_H^2 \leq 0. \tag{5.2b}$$

Adding (5.2a) and (5.2b) yields

$$(\partial_t AU^\mu + BU^\mu, G(U^\mu) - \bar{V}) \leq -\mu^{-1}\|\bar{U}^\mu - P_{\bar{K}}\bar{U}^\mu\|_H^2 \leq 0. \tag{5.3}$$

Choosing $\bar{V} = (0,0,0) \in \bar{K}$ and integrating from 0 to t, we obtain almost everywhere in $(0,T)$ the classical energy estimate

$$\begin{aligned} &\left[\int_\Omega \left(\frac{1}{2}|v^\mu|^2 + \frac{1}{2}\boldsymbol{\sigma}^\mu : N\boldsymbol{\sigma}^\mu + \mathbf{g}(\gamma^\mu)\right)(x,t)\,dx\right]_0^t \\ &\qquad \leq -\mu^{-1}\|G(U^\mu) - P_{\bar{K}}(G(U^\mu))\|_{L^2(0,T;H)}^2 \leq 0, \end{aligned} \tag{5.4}$$

where $\mathbf{g}(\gamma) \overset{\text{def}}{=} \int_0^\gamma g(c)dc$. Here, we have again used the boundary conditions and the antisymmetric nature of B.

From (5.4), even if we ignore Theorem 4.1, we can extract a subsequence—still denoted by U^μ—such that

$$U^\mu \rightharpoonup U \text{ in } L^\infty(0,T;H) \text{ weak-star}, \tag{5.5}$$

i.e.,

$$\begin{aligned} &v^\mu \rightharpoonup v \text{ in } L^\infty(0,T;(L^2(\Omega))^2) \text{ weak-star}, \\ &\boldsymbol{\sigma}^\mu \rightharpoonup \boldsymbol{\sigma} \text{ in } L^\infty(0,T;M_s(L^2(\Omega)))\text{weak-star}, \\ &\gamma^\mu \rightharpoonup \gamma \text{ in } L^\infty(0,T;L^2(\Omega)) \text{ weak-star}, \\ &g(\gamma^\mu) \rightharpoonup g^* \text{ in } L^\infty(0,T;L^2(\Omega)) \text{ weak-star}. \end{aligned} \tag{5.6}$$

We note that, in view of assumptions (H1) and (H2), $\mathbf{g}$ is a convex nonnegative function such that

$$\forall \gamma \in \mathbb{R}, \frac{\alpha}{2}\gamma^2 \le \mathbf{g}(\gamma) \le \frac{\beta}{2}\gamma^2 \quad \text{and} \quad (g(\gamma))^2 \le \beta^2\gamma^2 \le 2\frac{\beta^2}{\alpha}\mathbf{g}(\gamma). \tag{5.7}$$

On the other hand, from Theorem 4.1, we can extract a new subsequence such that

$$\begin{aligned}
&\mu^{-1}(P_{\bar{K}}\bar{U}^\mu - \bar{U}^\mu) \to \chi \quad \text{in } L^\infty(0,T;L^2(\Omega)) \text{ weak-star},\\
&\partial_t v^\mu = \operatorname{div}\boldsymbol{\sigma}^\mu \rightharpoonup \partial_t v = \operatorname{div}\boldsymbol{\sigma} \quad \text{in } L^\infty(0,T;(L^2(\Omega)^3) \text{ weak-star},\\
&\partial_t \boldsymbol{\sigma}^\mu = N^{-1}(\boldsymbol{\varepsilon}(v^\mu) + \mu^{-1}[P_{\bar{K}}\bar{U}^\mu - \bar{U}^\mu]_2)\\
&\qquad \rightharpoonup \partial_t\boldsymbol{\sigma} = N^{-1}(\boldsymbol{\varepsilon}(v) + [\chi]_2) \quad \text{in } L^\infty(0,T;M_s(L^2(\Omega))) \text{ weak-star},\\
&\partial_t\gamma^\mu = \mu^{-1}[P_{\bar{K}}\bar{U}^\mu - \bar{U}^\mu]_3 \rightharpoonup [\chi]_3 \quad \text{in} L^\infty(0,T;L^2(\Omega)) \text{ weak-star},\\
&\boldsymbol{\varepsilon}(v^\mu) \rightharpoonup \boldsymbol{\varepsilon}(v^\mu) \quad \text{in } L^\infty(0,T;M_s(L^2(\Omega))) \text{ weak-star}.
\end{aligned} \tag{5.8}$$

Here, for any $U = (v, \boldsymbol{\sigma}, \gamma)$ in H, we have defined

$$[U]_1 = v, \qquad [U]_2 = \boldsymbol{\sigma}, \qquad [U]_3 = \gamma.$$

Clearly χ satisfies

$$[\chi]_1 = 0.$$

We also note that the natural energy estimate (5.4) would only imply

$$\|(I - P_{\bar{K}})(\bar{U}^\mu)\|_{L^\infty(0,T;H)} = O(\mu^{1/2}) \tag{5.9}$$

in contrast with the much better estimate (4.22).

Step 2. Since g is convex,

$$g^* \ge g(\gamma) \quad \text{a.e. in } Q = \Omega x(0,T). \tag{5.10}$$

For the same reason

$$\begin{aligned}
&w^* \text{ limit } (|\boldsymbol{\sigma}^{D\mu}|) \ge |\boldsymbol{\sigma}^D| \quad \text{a.e. in } Q,\\
&w^* \text{ limit } (|v^\mu|^2) \ge |v|^2 \quad \text{a.e. in } Q,\\
&w^* \text{ limit } ((\boldsymbol{\sigma}^\mu; N\boldsymbol{\sigma}^\mu)) \ge (\boldsymbol{\sigma} : N\boldsymbol{\sigma}) \quad \text{a.e. in } Q,
\end{aligned} \tag{5.11}$$

and

$$P_{\bar{K}}\bar{U}^\mu = P_{\bar{K}}(G(U^\mu)) \to \zeta \in \bar{K} \quad \text{in } L^\infty(0,T;H) \text{ weak-star}. \tag{5.12}$$

But, from (4.22) or (5.9),

$$\bar{U}^\mu - P_{\bar{K}}\bar{U}^\mu \to 0 \quad \text{in } L^\infty(0,T;H) \text{ strongly}. \tag{5.13}$$

Therefore,

$$G(U^\mu) = \bar{U}^\mu \rightharpoonup \bar{U}^* = (v, \boldsymbol{\sigma}^D, g^*) = \zeta \in \bar{K} \text{ in } L^\infty(0,T;H) \text{ weak-star}$$

and by using (5.10) this implies

$$|\boldsymbol{\sigma}^D| + g(\gamma) \le |\boldsymbol{\sigma}^D| + g^* \le \sigma_y, \quad \text{i.e., } U \in K \text{ a.e. in } Q. \tag{5.14}$$

Combining (5.4), (5.6), (5.10), and (5.11), we obtain, since $\gamma(\cdot,0)\equiv 0$,

$$\int_\Omega \left(\frac{1}{2}|v|^2+\frac{1}{2}\boldsymbol{\sigma}:N\boldsymbol{\sigma}+g(\gamma)\right)(x,t)\,dx \le \int_\Omega \left(\frac{1}{2}|v_0|^2+\frac{1}{2}\boldsymbol{\sigma}_0:N\boldsymbol{\sigma}_0\right)(x)\,dx. \tag{5.15}$$

On the other hand, from (5.8), $U=(v,\boldsymbol{\sigma},\gamma)$ satisfies

$$\begin{cases} \partial_t v - \ \mathrm{div}(\boldsymbol{\sigma}) = [\chi]_1 = 0, \\ \partial_t N\boldsymbol{\sigma} - \boldsymbol{\varepsilon}(v) = [\chi]_2, \\ \partial_t \gamma = [\chi]_3, \end{cases} \tag{5.16}$$

with $[\chi]_2 \in L^\infty(0,T;M_s(L^2(\Omega))$ and $[\chi]_3 \in L^\infty(0,T;L^2(\Omega))$, and the boundary conditions

$$\boldsymbol{\sigma}\cdot n = 0 \quad \text{on } \partial\Omega_1, \tag{5.17}$$

$$v = 0 \quad \text{on } \partial\Omega_2. \tag{5.18}$$

Here, we have first noted that, due to the Korn inequality,

$$v^\mu \to v \quad \text{in } L^\infty(0,T;(H^1(\Omega))^3) \text{ weak-star} \tag{5.19}$$

and

$$\partial_t v^\mu \to \partial_t v \quad \text{in } L^\infty(0,T;(L^2(\Omega))^3) \text{ weak-star.} \tag{5.20}$$

Since the trace of a function in $H^1(\Omega)$ is well defined, and lies in the Sobolev space $H^{1/2}(\partial\Omega)$, (5.19) allows us to define $v_{/\partial\Omega_2}$ and therefore to justify (5.18).

Since $\mathrm{div}\boldsymbol{\sigma} \in L^\infty(0,T;(L^2(\Omega))^3)$, we can define $(\boldsymbol{\sigma}\cdot n)_{/\partial\Omega_1}$ with a similar argument and thus justify (5.17). The same arguments also give a sense to the initial data

$$(v,\boldsymbol{\sigma},\gamma)(x,0) \equiv (v_0,\boldsymbol{\sigma}_0,0)(x). \tag{5.21}$$

Step 3. Now, let us show that $U=(v,\boldsymbol{\sigma},\gamma)$ is a globally defined weak solution to $(\mathcal{P})$, i.e., satisfies (5.1). Since, by (5.8)

$$\partial_t AU^\mu \rightharpoonup \partial_t AU \quad \text{in } L^\infty(0,T;H) \text{ weak-star} \tag{5.22}$$

and

$$BU^\mu \rightharpoonup BU \quad \text{in } L^\infty(0,T;H) \text{ weak-star,} \tag{5.23}$$

we clearly have, for all $\bar{V} \in L^1(0,T;H)$,

$$\int_0^T (\partial_t AU^\mu + BU^\mu, \bar{V})_H\,dt \to \int_0^T (\partial_t AU + BU, \bar{V})_H\,dt. \tag{5.24}$$

On the other hand, multiplying (5.16) by $G(U)$, integrating, and using formulas (5.17), (5.18), we obtain the classical energy estimate

$$\begin{aligned} \int_0^T (\partial_t AU + BU, G(U))_H\,dt &= \left[\int_\Omega \left(\frac{1}{2}|v|^2+\frac{1}{2}\boldsymbol{\sigma}:N\boldsymbol{\sigma}+g(\gamma)\right)(x,t)\,dx\right]_{t=0}^{t=T} \\ &= \int_0^T (\chi, G(U))_H\,dt < +\infty. \end{aligned} \tag{5.25}$$

Now, we come back to (5.3), where $\bar{V} \in L^1(0,T;H)$ takes arbitrary values in $\bar{K}$, and we integrate with respect to t

$$\int_0^T (\partial_t AU^\mu + BU^\mu, G(U^\mu) - \bar{V})_H \, dt \leq 0. \tag{5.26}$$

We now return to (5.4) and (5.15). First, if we replace (0, t) by any time interval $(t', t), t' < t$, we see that for every μ the total energy

$$E(U^\mu)(t) \stackrel{\text{def}}{=} \int_\Omega E(U^\mu)(x,t)\, dx \stackrel{\text{def}}{=} \int_\Omega \left(\frac{1}{2}|v^\mu|^2 + \frac{1}{2}\boldsymbol{\sigma}^\mu : N\boldsymbol{\sigma}^\mu + g(\gamma^\mu) \right)(x,t)\, dx$$

is a nonincreasing function of time. Therefore, $E(U^\mu)$ (or $E(U)$) is defined for all t in $[0,T]$, except perhaps on a countable set of points of discontinuity. Moreover, the sequence $(E(U^\mu))$ is bounded in $L^\infty(0,T)$. On the other hand, $E(U)$ is a convex function of U. Therefore, the weak-star limits, defined by $U^\mu \rightharpoonup U, E(U^\mu) \rightharpoonup E^*$, satisfy

$$E^* \geq E(U).$$

Multiplying $E(U^\mu)(t)$ by an arbitrary nonnegative test function $\phi(t)$ and applying Fatou's lemma, we obtain

$$\begin{aligned} \int_0^T \Phi(t) E(U)(t)\, dt &\leq \int_0^T \int_\Omega E^*(x,t)\, dx\, dt = \overline{\lim}_{\mu\to 0} \int_0^T \Phi(t) E(U^\mu)(t)\, dt \\ &\leq \int_0^T \Phi(t) \overline{\lim}_{\mu\to 0} E(U^\mu)(t)\, dt. \end{aligned} \tag{5.27}$$

Therefore, we have almost everywhere in $(0,T)$

$$\left[\int_\Omega E(x,t)\, dx \right]_0^t \leq \overline{\lim}_{\mu\to 0} \left[\int_\Omega E^\mu(x,t)\, dx \right]_0^t. \tag{5.28}$$

In this formula we can choose $t = T$ (or at worst $t = T - 0$; see also Remark 5.1). Combining with (5.3), (5.24), (5.25), and (5.26), we obtain

$$\begin{aligned} \int_0^T (\partial_t AU + BU, G(U))_H\, dt &= \left[\int_\Omega E(x,t)\, dx \right]_0^T \\ &\leq \overline{\lim}_{\mu\to 0} \left[\int_\Omega E^\mu(x,t)\, dx \right]_0^T \\ &= \overline{\lim}_{\mu\to 0} \int_0^T (\partial_t AU^\mu + BU^\mu, G(U^\mu))_H\, dt \\ &\leq \overline{\lim}_{u\to 0} \int_0^T (\partial_t AU^\mu + BU^\mu, \bar{V})_H\, dt \\ &= \int_0^T (\partial_t AU + BU, \bar{V})_H\, dt, \end{aligned} \tag{5.29}$$

which is exactly the desired result.

Step 4. First, let us prove the strong convergence of v^μ and $\boldsymbol{\sigma}^\mu$ to v and $\boldsymbol{\sigma}$ in $L^p(0,T;(L^2(\Omega))^3)$ and $L^p(0,T;(M_s(\Omega)))$, respectively, for every $p < +\infty$. By definition of the projection on a convex set, and of the subdifferential ∂I_K, taking $\bar{V} = \bar{U}$ in (5.2a) and $\bar{V} = P_{\bar{K}}\bar{U}^\mu$ in (5.29), we have

$$\int_0^T (\partial_t AU^\mu + BU^\mu, P_{\bar{K}}\bar{U}^\mu - \bar{U})_H \, dt \leq 0 \tag{5.30}$$

and

$$\int_0^T (-\partial_t AU - BU, P_{\bar{K}}\bar{U}^\mu - \bar{U})_H \, dt \leq 0. \tag{5.31}$$

Hence

$$\int_0^T (\partial_t A(U^\mu - U) + B(U^\mu - U), P_{\bar{K}}\bar{U}^\mu - \bar{U})_H \, dt \leq 0. \tag{5.32}$$

Since

$$\partial_t A(U^\mu - U) + B(U^\mu - U) \to 0 \text{ in } L^\infty(0,T;(L^2(\Omega))^3) \text{ weak-star}$$

and

$$\bar{U}^\mu - P_{\bar{K}}\bar{U}^\mu \to 0 \text{ in } L^1(0,T;(L^2(\Omega))^3) \text{ strong},$$

$$\int_0^T (\partial_t A(U^\mu - U) + B(U^\mu - U), \bar{U}^\mu - P_{\bar{K}}\bar{U}^\mu)_H \, dt \to 0. \tag{5.33}$$

The same result is true for almost all t in $(0,T)$; see Remark 5.1.

$$\overline{\lim}_{\mu\to 0} \int_0^t (\partial_t A(U^\mu - U) + B(U^\mu - U), \bar{U}^\mu - \bar{U})_H \, d\tau \leq 0, \tag{5.34}$$

i.e.,

$$\begin{aligned} \overline{\lim}_{\mu\to 0} \Bigg[\int_\Omega \left(\frac{1}{2}|v^\mu - v|^2 + \frac{1}{2}(\boldsymbol{\sigma}^\mu - \boldsymbol{\sigma}) : N(\boldsymbol{\sigma}^\mu - \boldsymbol{\sigma}) \right) d\tau \Bigg]_0^t \\ + \int_0^t \int_\Omega (g(\gamma^\mu) - g(\gamma))(\partial_t\gamma^\mu - \partial_t\gamma) \, dx \, dt \leq 0. \end{aligned} \tag{5.35}$$

Define

$$A^\mu = (g(\gamma^\mu) - g(\gamma))(\partial_t\gamma^\mu - \partial_t\gamma)$$

and use the Taylor formula to compute

$$\begin{aligned} A^\mu &= \left(\int_0^1 g'(\gamma + h(\gamma^\mu - \gamma)) \, dh \right) (\gamma^\mu - \gamma)(\partial_t\gamma^\mu - \partial_t\gamma) \\ &= \partial_t \left(\left(\int_0^1 g'(\gamma + h(\gamma^\mu - \gamma)) \, dh \right) \frac{(\gamma^\mu - \gamma)^2}{2} \right) \\ &\quad - \frac{(\gamma^\mu - \gamma)^2}{2} \int_0^t g''(\gamma + h(\gamma^\mu - \gamma))(\partial_t\gamma + h(\partial_t\gamma^\mu - \partial_t\gamma)) \, dh. \end{aligned} \tag{5.36}$$

Now $\partial_t \gamma^\mu$ and $\partial_t \gamma$ are nonpositive, and g is convex. Therefore,

$$\int_0^t \int_\Omega A^\mu \, dx \, d\tau \geq \int_\Omega \left[\left(\int_0^1 g'(\gamma + h(\gamma^\mu - \gamma))(\gamma^\mu - \gamma) \, dh \right) \frac{1}{2} (\gamma^\mu - \gamma)(x,\tau) \, dx \right]_0^t$$

$$= \frac{1}{2} \left[\int_\Omega (g(\gamma^\mu) - g(\gamma))(\gamma^\mu - \gamma)(x,\tau) \, dx \right]_0^t, \tag{5.37}$$

which implies for almost all t in $(0,T)$:

$$\overline{\lim}_{\mu \to 0} \left[\int_\Omega \left(\frac{1}{2} |v^\mu - v|^2 + \frac{1}{2} (\boldsymbol{\sigma}^\mu - \boldsymbol{\sigma}) : N(\boldsymbol{\sigma}^\mu - \boldsymbol{\sigma}) \right. \right.$$

$$\left. \left. + (g(\gamma^\mu) - g(\gamma))(\gamma^\mu - \gamma) \right) dx \right]_0^t \leq 0. \tag{5.38}$$

Consequently, using the assumption (H2) on g,

$$\|U^\mu(t) - U(t)\|_H \to 0 \quad \text{a.e. in } (0,T). \tag{5.39}$$

Therefore, by the Lebesgue theorem, the subsequence (U^μ) converges strongly to U in the space $L^p(0,T;H), \forall p < +\infty$.

Finally, we can use the same method to compare two solutions U_1 and U_2 associated to the same initial data and boundary conditions. We obtain

$$\left[\int_\Omega \left(\frac{1}{2} |v_1 - v_2|^2 + \frac{1}{2} (\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2) : N(\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2) \right. \right.$$

$$\left. \left. + (g(\gamma_1) - g(\gamma_2))(\gamma_1 - \gamma_2) \right) (x,t) \, dx \right]_0^t \leq 0, \tag{5.40}$$

which guarantees the uniqueness of the solution U and therefore the convergence of the whole sequence (U^μ). Theorem (5.1) is thus entirely proved.

Remark 5.2. As we said in the introduction, the problem is not associated with a monotone operator. Nevertheless, it satisfies property (5.40), which is very close to monotonicity but which is only true for solutions to the evolution problem or more generally for functions U_1 such that $\partial_t \gamma_i \leq 0, i = 1,2$.

6. Appendix: Some basic facts from convex analysis. We first recall that the indicator function of a set K is defined by

$$I_K(x) = \begin{cases} 0 & \text{if } x \in K, \\ +\infty & \text{if } x \notin K. \end{cases}$$

If K is a closed convex set, then I_K is a lower semicontinuous convex function. If K is a closed convex set in a Hilbert space H, whose scalar product is denoted by $(,)$, the subdifferential of I_K, denoted by ∂I_K is defined by

$$\partial I_K(x) = \{z \in H \text{ s.t. } (z, y - x) \leq 0 \text{ for any } y \text{ in } K\}.$$

$\partial I_K(x)$ is a closed convex set, centered in 0, geometrically, given by

$$\partial I_K(x) = \begin{cases} \varnothing & \text{if } x \in K, \\ 0 & \text{if } x \in \text{Int}(K), \\ \text{the exterior normal cone of } K \text{at } x & \text{if } x \in \partial K, \end{cases}$$

where $\mathrm{Int}(K)$ and ∂K, respectively, denote the interior and the boundary of K. In the particular case where K is a smooth convex set, the exterior normal cone of K reduces to the exterior normal to K at x.

Remark. After completion of this work, we have learned that a similar problem—with a more general class of (vector-valued) hardening parameters—is considered in Lami Dozo and Muler [10]. In this paper, the solution satisfies a weak version of the constitutive relation, which would be the classical one if there were enough regularity to satisfy our formula (5.25). Here, we have obtained precisely such a regularity result, which in addition guarantees the uniqueness of the solution.

REFERENCES

[1] S. S. ANTMAN, *Nonlinear elastoplastic waves*, in Current Progress in Hyperbolic System: Riemann Problems and Computations, Contemp. Math., Amer. Math. Soc., Providence, RI, 100 (1989), pp. 27–54.

[2] J. AUBIN, *Un théorème de compacité*, C. R. A. S. Paris, 265 (1963), pp. 5042–5044.

[3] D. BLANCHARD, P. LE TALLEC, AND M. RAVACHOL, *Numerical analysis of evolution problems in nonlinear small stains elastoviscoplasticity*, Numer. Math, 55 (1989), pp. 177–195.

[4] H. BREZIS, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland Mathematics Studies, Amsterdam, 1973.

[5] J. L. CHABOCHE, AND J. LEMAITRE, *Mécanique des matériaux solides.* Dunod, Paris, 1985, pp. 197–203.

[6] G. DUVAUT AND J. L. LIONS, *Les inéquations en mécanique et en physique.* Dunod, Paris, 1972.

[7] I. EKELAND AND T. TEMAM, *Analyse convexe et problèmes variationnels.* Dunod, Paris, 1974.

[8] P. GERMAIN, *Cours de Mécanique des milieux continus.* Masson, Paris, 1973.

[9] P. LABORDE AND Q. S. NGUYEN, *Etude de l'équation d'évolution des systèmes dissipatifs standards.* Modèlisation Mathématique et Analyse Numérique, 24 (1990), pp. 67–84.

[10] E. LAMI DOZO AND N. MULER, *Dynamical problems for elastic plastic bodies with nonlinear hardening*, preprint.

[11] D. G. SCHAEFFER AND M. SHEARER, *Scale-invariant initial value problems in one dimensional dynamic elastoplasticity, with consequences for multidimensional nonassociative plasticity*, European J. Appl. Math., 3 (1992), pp. 225–254.

[12] ———, *The initial value problem for a system modelling unidirectional longitudinal elastic-plastic waves*, SIAM J. Math. Anal., 24 (1993), pp. 1111–1144.

[13] P. SUQUET, *Evolution problems for a class of dissipative materials.* Quart. Appl. Math., 38 (1982), pp. 391–414.

[14] R. TEMAM, *Problèmes mathématiques en plasticité.* Bordas, Paris, 1983.

[15] J. A. TRANGENSTEIN AND R. PEMBER, *The Riemann problem for longitudinal motion in an elastic-plastic bar.* SIAM J. Sci. Statist. Comput., 12 (1991), pp. 180–207.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 869–879, July 1995

005

ON COUPLED INTEGRAL H-LIKE EQUATIONS OF CHANDRASEKHAR*

JONQ JUANG†

Abstract. A recently proposed "simple transport model" equation with an "angular shift" α $(0 \le \alpha \le 1)$ leads to a coupled integral H-like equation of Chandrasekhar. Such coupled H-like equations can be treated in terms of a one-parameter $(k_1,\ 0 < k_1 < 1)$ family. From there an a priori bound can be obtained, which is independent of k_1, α, and c $(0 \le c \le 1)$. Here c denotes the average total number of particles emerging from a collision. Consequently, we conclude that positive solutions of such coupled integral H-like equations exist. Moreover, we show that such equations have a unique positive solution pair for $c = 0$ or $c = 1$ and $\alpha = 0$ or $\alpha = 1$, and that the equations have exactly two positive solution pairs for $0 < c < 1$ and $0 \le \alpha < 1$ or $c = 1$ and α sufficiently close to 1.

Key words. integral equation, H-like functions of Chandrasekhar, a priori bound, existence and multiplicity

AMS subject classifications. 45G10, 82C70, 85A25

1. Introduction.

In this work we study the coupled integral H-like equations of the form

$$H_1(\mu) = 1 + \frac{c}{2} H_1(\mu)(\mu + \alpha) \int_\alpha^1 \frac{H_2(\mu'')}{\mu + \mu''}\, d\mu'', \quad -\alpha \le \mu \le 1, \tag{1a}$$

and

$$H_2(\mu') = 1 + \frac{c}{2} H_2(\mu')(\mu' - \alpha) \int_{-\alpha}^1 \frac{H_1(\mu'')}{\mu' + \mu''}\, d\mu'', \quad \alpha \le \mu' \le 1. \tag{1b}$$

Here c denotes the average total number of particles emerging from a collision, which is assumed to be conservative, i.e., $c \le 1$, and α denotes an "angular shift" with $0 \le \alpha \le 1$. Equation (1) first appeared in [8], where it was derived from a "simple transport model" (see, e.g., [5], [8]) using Chandrasekhar's method of solution. For $\alpha = 0$, equation (1) reduces to Chandrasekhar's well-known integral equation. Various methods (see, e.g., [1]–[4], [6], [7], [10], [11]) have been applied to such equations. In summary, they have shown that Chandrasekhar's integral equation has one solution if $c = 1$ and at most two solutions if $c < 1$.

In this article, we first show that an a priori bound, which is independent of c and α, can be obtained by introducing a one-parameter $(k_1, 0 < k_1 < 1)$ family. Therefore, the degree theory is applied to show the existence of positive solutions. Second, the techniques used in [6], [10] are generalized to show that equation (1) has a unique positive solution pair for $c = 0$ or $c = 1$ and $\alpha = 0$ or $\alpha = 1$, and that equation (1) has exactly two positive solution pairs for $0 < c < 1$ and $0 \le \alpha < 1$ or $c = 1$ and α sufficiently close to 1. The above results are contained in §2.

We conclude this introductory section by noting that using the solutions obtained by equation (1), the simple transport model can then be treated as a "pure" initial

* Received by the editors June 15, 1993; accepted for publication (in revised form) December 15, 1993. This research was partially supported by the National Science Council of the Republic of China.

† Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan, Republic of China.

value problem. More precisely, consider the following simple transport model:

$$(\mu+\alpha)\frac{\partial\phi(x,\mu)}{\partial x}+\phi(x,\mu)=\frac{c}{2}\int_{-1}^{1}\phi(x,\mu')\,d\mu',\ 0\le x<\infty,\ |\mu|\le 1,$$
$$\phi(0,\mu)=f(\mu),\ 1\ge\mu>-\alpha.$$

Then, for $-1\le\mu\le-\alpha$, we have (see equations (3) and (12) of [8]) that

$$\phi(0,\mu)=\frac{c}{2}\int_{-\alpha}^{1}\frac{\mu'+\alpha}{\mu'-\mu}H_1(\mu')H_2(-\mu)f(\mu')\,d\mu'.$$

Such an approach provides an interesting and effective alternative for solving the simple transport model theorectically as well as numerically.

2. Main results.

Notation. Set

$$x=\frac{c}{2}\int_{-\alpha}^{1}H_1(\mu)\,d\mu,\quad y=\frac{c}{2}\int_{\alpha}^{1}H_2(\mu')\,d\mu',$$

$$a=\frac{c^2}{4}\int_{-\alpha}^{1}\int_{\alpha}^{1}H_1(\mu)H_2(\mu'')\frac{\mu''-\alpha}{\mu+\mu''}\,d\mu''\,d\mu,$$

and

$$b=\frac{c^2}{4}\int_{-\alpha}^{1}\int_{\alpha}^{1}H_1(\mu)H_2(\mu'')\frac{\mu+\alpha}{\mu+\mu''}\,d\mu''\,d\mu.$$

Note that $a+b=xy$. We begin by deriving some integral properties which a solution of (1) must satisfy.

LEMMA 1. *If H_1 and H_2 are solutions of* (1), *then the following holds:*

$$(1-x)(1-y)=1-c. \tag{2}$$

Proof. Multiplying equation (1) by $\frac{c}{2}$ and integrating equations (1a) and (1b) over the ranges of μ' and μ, respectively, we obtain

$$x=\frac{c}{2}(1+\alpha)+xy-a \tag{3a}$$

and

$$y=\frac{c}{2}(1-\alpha)+xy-b. \tag{3b}$$

Adding up (3a) and (3b) would yield the assertion of Lemma 1.

Remark. For $\alpha=0$, (2) reduces to some well-known expressions concerning the properties of H equations (see, e.g., [3, pp. 106–107]).

For $\alpha\neq 1$, we see immediately that if H_1 and H_2 are positive solutions of (1), then there must exist two positive numbers k_1 and k_2, where $0<k_1, k_2<1$ and $k_1+k_2=1$, so that

$$a=k_1xy \tag{4a}$$

and

$$b=k_2xy. \tag{4b}$$

It then follows from (2), (3), and (4) that the following holds:

$$\text{(5a)}\quad x = \frac{1-\frac{c}{2}(1-\alpha)+k_1c \pm \sqrt{[1-\frac{c}{2}(1-\alpha)+k_1c]^2 - 2k_1(1+\alpha)c}}{2k_1} := a_1 \pm b_1,$$

$$\text{(5b)}\quad y = \frac{1-\frac{c}{2}(1+\alpha)+k_2c \pm \sqrt{[1-\frac{c}{2}(1+\alpha)+k_2c]^2 - 2k_2(1-\alpha)c}}{2k_2} := a_2 \pm b_2.$$

Since k_1 and k_2 are to be treated as real parameters, necessary conditions for (5) to be meaningful are that both $[1-\frac{c}{2}(1-\alpha)+k_1c]^2 - 2k_1c(1+\alpha)$ and $[1-\frac{c}{2}(1+\alpha)+k_2c]^2 - 2k_2c(1-\alpha)$ are nonnegative. However, these are so if $0 \le \alpha \le 1$ and $0 \le c \le 1$. To see this, we note that, for $c \neq 0$, $f_1(k_1) := [1-\frac{c}{2}(1-\alpha)+k_1c]^2 - 2k_1c(1+\alpha)$ has a minimim $(1+\alpha)(1-\alpha)(1-c)$, which is nonnegative whenever $0 \le \alpha \le 1$ and $0 \le c \le 1$.

We denote by S the feasible region $\{(k,c,\alpha) : 0 < k < 1, 0 \le c \le 1 \text{ and } 0 \le \alpha \le 1\}$ for the solution of (1). The cross-section $\{(k,c,\alpha) : 0 < k < 1, 0 \le c \le 1 \text{ and } \alpha \text{ is fixed}\}$ of S will be denoted by S_α. The properties and signs of $1-x$ and $1-y$ will be examined in the next lemmas.

LEMMA 2. (i) $1-a_1+b_1 \ge 0$ *and* $1-a_1-b_1 \le 0$ *for all* $(k_1,c,\alpha) \in S$.

(ii) $1-a_2+b_2 \ge 0$ *and* $1-a_2-b_2 \le 0$ *for all* $(k_2,c,\alpha) \in S$.

(iii) *For each fixed* α, *where* $0 \le \alpha < 1$, *we have that* $1-a_1+b_1$ *and* $1-a_2+b_2$, *considered as functions from* $S_\alpha \to R$, *can be continuously extended to* $\bar{S}_\alpha$.

(iv) *Let* c *be sufficiently small, say* $0 \le c \le \frac{1}{8}$. *Then* $1-a_1+b_1 \ge \frac{1}{2}$ *and* $1-a_2+b_2 \ge \frac{6}{7}$ *for all* k_1 *and* k_2, $0 < k_1, k_2 < 1$, *and all* α, $0 \le \alpha \le 1$.

Proof. Since the computation leading to (i) and (ii) is similar, we shall only illustrate (i). To see (i), it suffices to show that $b_1^2 \ge (1-a_1)^2$, or equivalently

$$\left[1-\frac{c}{2}(1-\alpha)+k_1c\right]^2 - 2k_1c(1+\alpha) - \left[(2k_1-1)\left(1-\frac{c}{2}\right) - \frac{c\alpha}{2}\right]^2 \ge 0.$$

Since the left-hand side of the inequality is equal to $4(1-k_1)(k_1)(1-c)$, the assertion of Lemma 2(i) thus follows. To prove (iii), we note that

$$a_1 - b_1 = \frac{(1+\alpha)c}{1-\frac{c}{2}(1-\alpha)+k_1c+\sqrt{[1-\frac{c}{2}(1-\alpha)+k_1c]^2 - 2k_1c(1+\alpha)}} := \frac{(1+\alpha)c}{g_1(k_1,c,\alpha)}$$

and

$$a_2 - b_2 = \frac{(1-\alpha)c}{1-\frac{c}{2}(1+\alpha)+k_2c+\sqrt{[1-\frac{c}{2}(1+\alpha)+k_2c]^2 - 2k_2c(1-\alpha)}} := \frac{(1-\alpha)c}{g_2(k_2,c,\alpha)}.$$

Since $g_1(k_1,c,\alpha) \ge \frac{1}{2}$ for all $(k_1,c,\alpha) \in \bar{S}$, we conclude that $a_1 - b_1$, and hence $1-a_1+b_1$, can be continuously extended to $\bar{S}$. Now, if α is fixed as assumed, then $g_2(k_2,c,\alpha) \ge \frac{1}{2}(1-\alpha) > 0$ for all (k_2,c). Therefore, for each fixed α, $1-a_2+b_2$ can be continuously extended to $\bar{S}_\alpha$.

To prove (iv), we see that if $0 \le c \le \frac{1}{8}$, then $a_1 - b_1 = \frac{(1+\alpha)c}{g_1(k_1,c,\alpha)} \le 2c(1+\alpha) \le \frac{1}{2}$, for all k_1 and α. Thus, $1-a_1+b_1 \ge \frac{1}{2}$ as asserted. Similarly, we have

$$a_2 - b_2 = \frac{(1-\alpha)c}{g_2(k_2,c,\alpha)} \le \frac{c}{1-\frac{c}{2}(1+\alpha)} \le \frac{c}{1-c} \le \frac{1}{7}.$$

Therefore, $1-a_2+b_2 \ge \frac{6}{7}$ as asserted.

Remarks. The function $1-a_1+b_1$, as indicated in the proof, can be continuously extended to $\bar{S}$. However, the same assertion fails for $1-a_2+b_2$. To see this, we note that if $\alpha=1$, then $a_2-b_2=0$ for all k_2 and c. However, if $c=1$ then $a_2-b_2=1$ for any $\alpha(\neq 1)$ and $k_2 \leq \frac{1}{2}(1-\alpha)$.

In view of (2), the case $c=1$ shall be further studied.

LEMMA 3. (i) $1-a_1+b_1=0$ *if and only if* $\frac{1}{2}(1+\alpha) \geq k_1$ *and* $c=1$. *Moreover,* $1-a_1-b_1=0$ *if and only if* $\frac{1}{2}(1+\alpha) \leq k_1$ *and* $c=1$.

(ii) $1-a_2+b_2=0$ *if and only if* $k_1 \geq \frac{1}{2}(1+\alpha)$ *and* $c=1$. *Furthermore,* $1-a_2-b_2=0$ *if and only if* $k_1 \leq \frac{1}{2}(1+\alpha)$ *and* $c=1$.

(iii) *If* $\frac{1}{2}(1+\alpha)<k_1$ *and* $c=1$, *then* $1-a_1+b_1=\frac{2k_1-1-\alpha}{2k_1}$. *Moreover, if* $\frac{1}{2}(1+\alpha)>k_1$ *and* $c=1$, *then* $1-a_1-b_1=\frac{2k_1-1-\alpha}{2k_1}$.

(iv) *If* $k_1<\frac{1}{2}(1+\alpha)$ *and* $c=1$, *then* $1-a_2+b_2=\frac{1+\alpha-2k_1}{2(1-k_1)}$. *Furthermore, if* $k_1>\frac{1}{2}(1+\alpha)$ *and* $c=1$, *then* $1-a_2-b_2=\frac{1+\alpha-2k_1}{2(1-k_1)}$.

Proof. The necessary parts of Lemma 3(i) follow from (2) and some simple algebra. The remainder of the proof is trivial and thus omitted.

Some simple algebra would yield the following equivalent formulation of (1).

LEMMA 4. *The functions* H_1 *and* H_2 *satisfy, respectively,*

$$[H_1(\mu)]^{-1}=(1-y)+\frac{c}{2}\int_\alpha^1 \frac{\mu''-\alpha}{\mu+\mu''}H_2(\mu'')\,d\mu'' \tag{6a}$$

and

$$[H_2(\mu')]^{-1}=(1-x)+\frac{c}{2}\int_{-\alpha}^1 \frac{\mu''+\alpha}{\mu'+\mu''}H_1(\mu'')\,d\mu'' \tag{6b}$$

if and only if H_1 *and* H_2 *satisfy* (1a) *and* (1b), *respectively.*

In view of (2), we see that if H_1 and H_2 are solutions of (1), then either

$$1-x\geq 0 \text{ and } 1-y\geq 0 \tag{7a}$$

or

$$1-x\leq 0 \text{ and } 1-y\leq 0. \tag{7b}$$

Let $C[-\alpha,1]\times C[\alpha,1]$ be the Banach space of pairs of bounded real-valued continuous functions with sup norm. That is, if $(h_1, h_2)\in C[-\alpha,1]\times C[\alpha,1]$, then

$$\| (h_1, h_2) \|_\infty := \max\left\{ \max_{-\alpha\leq\mu\leq 1} | h_1(\mu) | := \| h_1 \|_\infty, \max_{\alpha\leq\mu\leq 1} | h_2(\mu') | := \| h_2 \|_\infty \right\}.$$

In preparation for the use of a homotopy invariance argument define, for $(K_1, K_2)\in C[-\alpha,1]\times C[\alpha,1]$,

$$\psi_{1,c}(K_2(\mu))=(1-y)+\frac{c}{2}\int_\alpha^1 \frac{\mu''-\alpha}{\mu+\mu''}\frac{1}{K_2(\mu'')}\,d\mu'', \tag{8a}$$

$$\psi_{2,c}(K_1(\mu'))=(1-x)+\frac{c}{2}\int_{-\alpha}^1 \frac{\mu''+\alpha}{\mu'+\mu''}\frac{1}{K_1(\mu'')}\,d\mu'', \tag{8b}$$

$$\psi_c(K_1(\mu), K_2(\mu'))=(\psi_{1,c}(K_2(\mu))), \psi_{2,c}(K_1(\mu')). \tag{8c}$$

An a priori bound, which is independent of k_1 and c, is obtained in the following lemma.

LEMMA 5. *Let K_1 and K_2 be any positive continuous solutions of $(K_1, K_2) = \psi_c(K_1, K_2)$ satisfying* (7a). *Then there is an $m > 0$ (independent of c and α) such that $K_1(\mu) \geq m$ and $K_2(\mu') \geq m$ for all $(\mu, \mu') \in [-\alpha, 1] \times [\alpha, 1]$, all $0 \leq \alpha \leq 1$, and all $0 \leq c \leq 1$.*

Proof. Clearly, $H_1 = \frac{1}{K_1}$ and $H_2 = \frac{1}{K_2}$ are positive solutions of (1). Consequently, $1 \geq K_1(\mu)$ and $1 \geq K_2(\mu')$ for all $[\mu, \mu') \in [-\alpha, 1] \times [\alpha, 1]$. Therefore,

$$K_1(\mu) \geq 1 - y + \frac{c}{2} \int_\alpha^1 \frac{\mu'' - \alpha}{\mu'' + 1}\, d\mu'' := 1 - y + g_1(c, \alpha)$$

and

$$K_2(\mu') \geq 1 - x + \frac{c}{2} \int_{-\alpha}^1 \frac{\mu'' + \alpha}{\mu'' + 1}\, d\mu'' := 1 - x + g_2(c, \alpha).$$

Since $1-y \geq 0$ and $1-x \geq 0$, there must exist positive constants k_1 and k_2, $k_1+k_2 = 1$, such that $1 - x = 1 - a_1 + b_1$ and $1 - y = 1 - a_2 + b_2$, where $a_1 - b_1$ and $a_2 - b_2$ are defined as in (5). Now, via Lemma 2(iv), $1 - x \geq \frac{1}{2}$ for $0 \leq c \leq \frac{1}{8}$. Since, for fixed c, $g_2(c, \alpha)$ is an increasing function (in α), we have that $\int_{-\alpha}^1 \frac{\mu''+\alpha}{\mu''+1}\, d\mu'' \geq 1 - \ell n 2$. Consequently,

$$K_2 \geq \min\left\{\frac{1}{2}, \frac{1}{16}(1 - \ell n 2)\right\} = \frac{1}{16}(1 - \ell n 2) := m_2$$

for all $(\mu, \mu') \in [-\alpha, 1] \times [\alpha, 1]$, all $0 \leq c \leq 1$, and all $0 \leq \alpha \leq 1$. On the other hand,

$$y = \frac{c}{2} \int_\alpha^1 \frac{1}{K_2(\mu')}\, d\mu' \leq \frac{c(1-\alpha)}{2m_2},$$

and so $1 - y \geq 1 - \frac{c(1-\alpha)}{2m_2}$. Hence, if $0 \leq c \leq m_2$ or $\alpha \geq 1 - m_2$, then $1 - y \geq \frac{1}{2}$. However, if $1 \geq c \geq m_2$ and $0 \leq \alpha \leq 1 - m_2$ then

$$g_1(c, \alpha) \geq \frac{m_2}{2} \int_{1-m_2}^1 \frac{\mu'' - (1 - m_2)}{\mu'' + 1}\, d\mu'' := \bar{m}_1 > 0.$$

Consequently, $K_1(\mu) \geq \min\{\frac{1}{2}, \bar{m}_1\} := m_1$ as asserted. The assertion of the lemma now follows by choosing $m = \min\{m_1, m_2\}$.

Remark. The lower bound for K_2 is not sharp. A better bound can be obtained. To see this, let c be such that $0 \leq c \leq \frac{2}{9-\ell n 2}$, then $a_1 - b_1 \leq 2c(1 + \alpha) \leq \frac{8}{9 \ell n 2}$. Thus, $1 - a_1 + b_1 \geq \frac{1-\ell n 2}{9-\ell n 2}$ for $0 \leq c \leq \frac{2}{9-\ell n 2}$. Hence,

$$K_2(\mu') \geq 1 - x + \frac{c}{2}(1 - \ell n 2) \geq \min\left\{\frac{1 - \ell n 2}{9 - \ell n 2}, \frac{1 - \ell n 2}{9 - \ell n 2}\right\} = \frac{1 - \ell n 2}{9 - \ell n 2}.$$

THEOREM 1. *For each α and c, where $0 \leq \alpha < 1$ and $0 \leq c \leq 1$, ψ_c has a fixed point satisfying* (7a).

Proof. Note, via Lemma 2(iii), that there exists a positive constant $\tilde{m}$ such that

$$\max\left\{\max_{(k_1, c, \alpha) \in \bar{S}_\alpha} (1 - a_1 + b_1), \max_{(k_2, c, \alpha) \in \bar{S}_\alpha} (1 - a_2 + b_2)\right\} \leq \tilde{m}.$$

Choose $a = \min(\frac{1}{2}, \frac{m}{2})$ and $b = \frac{1}{m} + \tilde{m} + 1$, where m is chosen as in Lemma 5. Set $D = \{(K_1, K_2) \in C[-\alpha, 1] \times C[\alpha, 1] : a < K_1(\mu), K_2(\mu') < b \text{ for all } (\mu, \mu') \in [-\alpha, 1] \times [\alpha, 1]\}$. Clearly, D is a nonempty bounded open subset of $C[-\alpha, 1] \times C[\alpha, 1]$, and $\psi_c : \bar{D} \to C[-\alpha, 1] \times C[\alpha, 1]$ is compact. Next, we show that if $(K_1, K_2) =$

$\psi_c(K_1, K_2)$ for $(K_1, K_2) \in \bar{D}$, then $(K_1, K_2) \in D$. To prove this note first that from the a priori bound for (K_1, K_2), we see that $K_1(\mu)$, $K_2(\mu') \geq m > a$ for all μ, μ'. Second,

$$\| (K_1, K_2) \|_\infty = \| \psi_c(K_1, K_2) \|_\infty \leq \tilde{m} + \frac{c(1+\alpha)}{2m} < b.$$

Thus $u \in D$ as asserted. The preparations for the use of degree are now complete. Consider the homotopy $I - \psi_c$. By homotopy invariance (see, e.g., Theorem 13.6 of [9]), since $(1,1) \in D$,

$$d(I - \psi_c, (0,0), D) = d(I - \psi_0, (0,0), D) = d(I, (1,1), D) = 1.$$

Therefore, the existence of equation (1) now follows from the Leray–Schauder fixed point theorem.

To show the uniqueness of equation (1) satisfying (7a), we need the following lemma.

LEMMA 6. *Equation* (1) *has minimal positive solutions* $H_{1,\min}(\mu)$ *and* $H_{2,\min}(\mu')$ *in the following sense : if* $H_1(\mu)$ *and* $H_2(\mu')$ *are positive solutions of* (1), *then* $H_{1,\min}(\mu) \leq H_1(\mu)$ *and* $H_{2,\min}(\mu') \leq H_2(\mu')$ *for all* μ, μ'.

Proof. Consider the two iterates $\{H_1^{(p)}\}$ and $\{H_2^{(p)}\}$ defined as follows:

$$H_1^{(1)}(\mu) = 1, \tag{9a}$$

$$H_2^{(1)}(\mu') = 1 \text{ for all } \mu, \mu', \tag{9b}$$

$$H_1^{(p+1)}(\mu) = 1 + \frac{c}{2} H_1^{(p)}(\mu)(\mu + \alpha) \int_\alpha^1 \frac{H_2^{(p)}(\mu'')}{\mu + \mu''} \, d\mu'', \tag{9c}$$

and

$$H_2^{(p+1)}(\mu') = 1 + \frac{c}{2} H_2^{(p)}(\mu')(\mu' - \alpha) \int_{-\alpha}^1 \frac{H_1^{(p)}(\mu'')}{\mu' + \mu''} \, d\mu''. \tag{9d}$$

Clearly, for each μ and μ', $\{H_1^{(p)}(\mu)\}$ and $\{H_2^{(p)}(\mu')\}$ are both increasing sequences. It follows from Theorem 1 that equation (1) has positive solutions, say $H_1(\mu)$ and $H_2(\mu')$. Since $H_1(\mu) \geq 1$ and $H_2(\mu') \geq 1$ for all μ, μ', an easy induction would yield $H_1^{(p)}(\mu) \leq H_1(\mu)$ and $H_2^{(p)}(\mu') \leq H_1(\mu')$ for all μ, μ' and all p. Hence, the sequences $\{H_1^{(p)}(\mu)\}$ and $\{H_2^{(p)}(\mu')\}$, respectively, converge upward to two limits, say $\bar{H}_1(\mu)$ and $\bar{H}_2(\mu')$. It then follows from the monotone convergence theorem that $\bar{H}_1$ and $\bar{H}_2$ solve equation (1), and that $\bar{H}_1(\mu) \leq H_1(\mu)$ and $\bar{H}_2(\mu') \leq H_2(\mu')$ for all μ, μ'. The proof of the lemma is thus complete.

THEOREM 2. *For* $c = 0$ *or* $\alpha = 1$, *equation* (1) *has unique solutions. Furthermore, for* $0 < c < 1$, *equation* (1) *has unique solutions* H_1 *and* H_2 *satisfying* (7a).

Proof. The uniqueness for $c = 0$ or $\alpha = 1$ is trivial. For $0 < c < 1$, and $1 - x > 0$ and $1 - y > 0$, we have that

$$1 - x_{\min} := 1 - \frac{c}{2} \int_{-\alpha}^1 H_{1,\min}(\mu) d\mu \geq 1 - x > 0,$$

$$1 - y_{\min} := 1 - \frac{c}{2} \int_{\alpha}^1 H_{2,\min}(\mu') d\mu' \geq 1 - y > 0,$$

and

$$[H_2(\mu')]^{-1} = \frac{1-c}{1-y} + \frac{c}{2}\int_{-\alpha}^{1} \frac{\mu''+\alpha}{\mu'+\mu''} H_1(\mu'')\,d\mu''$$

$$\geq \frac{1-c}{1-y_{\min}} + \frac{c}{2}\int_{-\alpha}^{1} \frac{\mu''+\alpha}{\mu'+\mu''} H_{1,\min}(\mu'')\,d\mu''$$

$$\geq [H_{2,\min}(\mu')]^{-1}.$$

Therefore, $H_2 = H_{2,\min}$, and hence $H_1 = H_{1,\min}$, and the lemma is proved.

Our final result is concerned with the number of positive solutions for equation (1). The techniques for proving this result are motivated by those of Leggett [10]. To this end, we first prove the following lemma.

LEMMA 7. *Let $0 < c < 1$ and $0 \leq \alpha < 1$, and let (H_1, H_2) and $(\bar{H}_1, \bar{H}_2)$ be positive solutions pairs of equation* (1) *satisfying* (7a) *and* (7b), *respectively. Then the following holds:*

(i) *There exist, respectively, two positive constants k_1 and k_2, where $0 < k_1 < \frac{1}{1+\alpha}$ and $0 < k_2 < \frac{1}{1-\alpha}$, such that*

$$\frac{c}{2}\int_{-\alpha}^{1} \frac{H_1(\mu'')}{1-k_1(\mu''+\alpha)}\,d\mu'' = 1 \tag{10a}$$

and

$$\frac{c}{2}\int_{\alpha}^{1} \frac{H_2(\mu'')}{1-k_2(\mu''-\alpha)}\,d\mu'' = 1. \tag{10b}$$

Furthermore, such choices of k_1 and k_2 are unique.

(ii) *There exist, respectively, two positive constants $\bar{k}_1$ and $\bar{k}_2$, where $0 < \bar{k}_1 < \frac{1}{1+\alpha}$ and $0 < \bar{k}_2 < \frac{1}{1-\alpha}$, such that*

$$\frac{c}{2}\int_{-\alpha}^{1} \frac{\bar{H}_1(\mu'')}{1+\bar{k}_2(\mu''+\alpha)}\,d\mu'' = 1 \tag{10c}$$

and

$$\frac{c}{2}\int_{\alpha}^{1} \frac{\bar{H}_2(\mu'')}{1+\bar{k}_1(\mu''-\alpha)}\,d\mu'' = 1. \tag{10d}$$

Moreover, such choices of $\bar{k}_1$ and $\bar{k}_2$ are unique.

Proof. Since the analysis leading to (10a), (10b), (10c), and (10d) is similar, we illustrate only (10b) and (10c). Define the function $T : (0, \frac{1}{1-\alpha}) \to R$ by

$$T(k) = \frac{c}{2}\int_{\alpha}^{1} \frac{H_2(\mu'')}{1-k(\mu''-\alpha)}\,d\mu''.$$

Then

$$\lim_{k\to(\frac{1}{1-\alpha})^-} T(k) = \frac{c}{2}\int_{\alpha}^{1} \frac{(1-\alpha)H_2(\mu'')}{1-\mu''}\,d\mu'', \tag{11}$$

since $(1-k(\mu''-\alpha))^{-1}$ increases monotonically with k, $0 < k < \frac{1}{1-\alpha}$. Note that the improper integral in (11) diverges to $+\infty$. Since $T(0) = \frac{c}{2}\int_{\alpha}^{1} H_2(\mu'')\,d\mu'' < 1$, and

since $T(k)$ is strictly increasing with $T(\frac{1}{1-\alpha}) = +\infty$, there exists a unique $k_2 \in (0, \frac{1}{1-\alpha})$ for which (10b) holds. Now suppose that $\bar{H}_1$ and $\bar{H}_2$ satisfy (1) and (7b). Then $\frac{c}{2}\int_{-\alpha}^{1} \bar{H}_1(\mu'')\,d\mu'' > 1$, and

$$\frac{c}{2}\int_{-\alpha}^{1} \frac{\bar{H}_1(\mu'')}{1+\frac{1}{1-\alpha}(\mu''+\alpha)}\,d\mu'' = \frac{c}{2}\int_{-\alpha}^{1} \frac{1-\alpha}{1+\mu''}\bar{H}_1(\mu'')\,d\mu''$$

$$= 1 - [\bar{H}_2(1)]^{-1} < 1.$$

Therefore, there exists a unique $\bar{k}_2$, $0 < \bar{k}_2 < \frac{1}{1-\alpha}$, such that (10c) holds.

THEOREM 3. *Equation* (1) *has exactly two positive solutions if* $0 < c < 1$ *and* $0 \le \alpha < 1$.

Proof. Let H_1 and H_2 be positive solutions of (1) satisfying (7a). Define

$$\bar{H}_1(\mu) = \frac{1+k_2\mu+k_2\alpha}{1-k_1\mu-k_1\alpha}H_1(\mu) \tag{12a}$$

and

$$\bar{H}_2(\mu'') = \frac{1+k_1\mu''-k_1\alpha}{1-k_2\mu''+k_2\alpha}H_2(\mu''). \tag{12b}$$

Here k_1 and k_2 are chosen as in Lemma 7. Now, using (10b), we find

$$\frac{c}{2}\int_{\alpha}^{1} \frac{\mu+\alpha}{\mu+\mu''}\cdot\bar{H}_2(\mu'')\,d\mu'' = \frac{c}{2}\int_{\alpha}^{1} \frac{(\mu+\alpha)(1+k_1\mu''-k_1\alpha)}{(\mu+\mu'')(1-k_2\mu''+k_2\alpha)}H_2(\mu'')\,d\mu''$$

$$= \frac{1-k_1\mu-k_1\alpha}{1+k_2\mu+k_2\alpha}\left(\frac{c}{2}\right)\int_{\alpha}^{1}\frac{\mu+\alpha}{\mu+\mu''}H_2(\mu'')\,d\mu''$$

$$+ \frac{(k_1+k_2)(\mu+\alpha)}{1+k_2\mu+k_2\alpha}\left(\frac{c}{2}\right)\int_{\alpha}^{1}\frac{H_2(\mu'')}{1-k_2\mu''+k_2\alpha}\,d\mu''$$

$$= \frac{1-k_1\mu-k_1\alpha}{1+k_2\mu+k_2\alpha}\left[1-\frac{1}{H_1(\mu)}\right] + \frac{(k_1+k_2)(\mu+\alpha)}{1+k_2\mu+k_2\alpha}$$

$$= 1 - \frac{1}{\bar{H}_1(\mu)}.$$

A similar computation would yield that

$$\frac{c}{2}\int_{-\alpha}^{1}\frac{\mu'-\alpha}{\mu'+\mu''}\bar{H}_1(\mu'')\,d\mu'' = 1 - \frac{1}{\bar{H}_2(\mu')}.$$

That is, $\bar{H}_1$ and $\bar{H}_2$ satisfy equation (1). Hence, $\bar{H}_1$ and $\bar{H}_2$ must satisfy either (7a) or (7b). Since H_1 and H_2 are the unique positive solutions of (1) satisfying (7a), and since $\bar{H}_1(\mu) > H_1(\mu)$, $\bar{H}_2(\mu') > H_2(\mu')$ for almost all μ, μ', $\bar{H}_1$ and $\bar{H}_2$ must satisfy (7b). Thus, we have shown that equation (1) has at least two positive solutions when c and α are as assumed. It remains to show that such an equation has at most two

solutions. To this end, we suppose that $\bar{H}_1$ and $\bar{H}_2$ are positive solutions satisfying (1) and (7b). Define

$$H_1(\mu) = \frac{1 - \bar{k}_1\mu - \bar{k}_1\alpha}{1 + \bar{k}_2\mu + \bar{k}_2\alpha}\bar{H}_1(\mu) \tag{13a}$$

and

$$H_2(\mu'') = \frac{1 - \bar{k}_2\mu'' + \bar{k}_2\alpha}{1 + \bar{k}_1\mu'' - \bar{k}_1\alpha}\bar{H}_2(\mu''). \tag{13b}$$

Here $\bar{k}_1$ and $\bar{k}_2$ are chosen as in Lemma 7. Now, using (10c), we obtain

$$\begin{aligned}
\frac{c}{2}\int_{-\alpha}^{1} \frac{\mu' - \alpha}{\mu' + \mu''} \cdot H_1(\mu'')\, d\mu'' &= \frac{c}{2}\int_{-\alpha}^{1} \frac{(\mu' - \alpha)(1 - \bar{k}_1\mu'' - \bar{k}_1\alpha)}{(\mu' + \mu'')(1 + \bar{k}_2\mu'' + \bar{k}_2\alpha)}\bar{H}_1(\mu'')\, d\mu'' \\
&= \frac{1 + \bar{k}_1\mu' - \bar{k}_1\alpha}{1 - \bar{k}_2\mu' + \bar{k}_2\alpha}\left(\frac{c}{2}\right)\int_{-\alpha}^{1} \frac{\mu' - \alpha}{\mu' + \mu''}\bar{H}_1(\mu'')\, d\mu'' \\
&\quad - \frac{(\bar{k}_1 + \bar{k}_2)(\mu' - \alpha)}{1 - \bar{k}_2\mu' + \bar{k}_2\alpha}\left(\frac{c}{2}\right)\int_{-\alpha}^{1} \frac{\bar{H}_1(\mu'')}{1 + \bar{k}_2\mu'' + \bar{k}_2\alpha}\, d\mu'' \\
&= \frac{1 + \bar{k}_1\mu' - \bar{k}_1\alpha}{1 - \bar{k}_2\mu' + \bar{k}_2\alpha}\left[1 - \frac{1}{\bar{H}_2(\mu')}\right] - \frac{(\bar{k}_1 + \bar{k}_2)(\mu' - \alpha)}{1 - \bar{k}_2\mu' + \bar{k}_2\alpha} \\
&= 1 - \frac{1}{\bar{H}_2(\mu')}.
\end{aligned}$$

Similarly, we obtain that

$$\frac{c}{2}\int_{\alpha}^{1} \frac{\mu + \alpha}{\mu + \mu''}H_2(\mu'')\, d\mu'' = 1 - \frac{1}{H_1(\mu)}.$$

Therefore, H_1 and H_2 satisfy equation (1). It follows from (13) and (10c), (10d) that $\frac{c}{2}\int_{-\alpha}^{1} H_1(\mu)\, d\mu < 1$ and $\frac{c}{2}\int_{\alpha}^{1} H_2(\mu')\, d\mu' < 1$; i.e., H_1 and H_2 satisfy (7a). Since the solutions of (1) satisfying (7a) are unique, we conclude that the solutions of (1) satisfying (7b) are also unique, and the theorem is proved.

THEOREM 4. *Let $c = 1$ and let α be sufficiently close to 1. Then equation (1) has exactly two positive solutions.*

Proof. Let H_1 and H_2 be solutions of equation (1) satisfying (7a). It follows from Lemma 5 that y must approach zero as α approaches 1 from the left. Hence if $c = 1$ and α is chosen to be sufficiently close to 1, then $x = 1$ and $y < 1$. Define $\bar{H}_1$ and $\bar{H}_2$ as follows:

$$\bar{H}_1(\mu) = (1 + k_2\mu + k_2\alpha)H_1(\mu)$$

and

$$\bar{H}_2(\mu'') = \frac{H_2(\mu'')}{1 - k_2\mu'' + k_2\alpha},$$

where k_2 is uniquely satisfied by (10b). Using a procedure similar to the proof of Theorem 3, it follows that $\bar{H}_1$ and $\bar{H}_2$ are positive solutions of (1) satisfying (7b). Since $H_1 \neq \bar{H}_1$ and $H_2 \neq \bar{H}_2$, it remains to show that such an equation has at most

two positive solutions. Suppose $\bar{H}_1$ and $\bar{H}_2$ are positive solutions of (1) satisfying (7b). Then either

$$\bar{x} := \frac{1}{2}\int_{-\alpha}^{1} \bar{H}_1(\mu)d\mu > 1 \text{ and } \bar{y} := \frac{1}{2}\int_{\alpha}^{1} \bar{H}_2(\mu')d\mu' = 1, \tag{14a}$$

or

$$\bar{x} = 1 \text{ and } \bar{y} > 1, \tag{14b}$$

or

$$\bar{x} = 1 \text{ and } \bar{y} = 1. \tag{14c}$$

If (14c) held, then $\bar{H}_1$ and $\bar{H}_2$ would also satisfy (7a), and hence $\bar{y} \to 0$ as $\alpha \to 1$, a contradication. Thus, (14c) should be ruled out. If (14b) were the case, then H_1 and H_2, defined as in (13a) and (13b), respectively, with $\bar{k}_1 = 0$ and k_2 satisfying (10c), were positive solutions of (1) satisfying (7a). Since $H_1 \leq \bar{H}_1$ and $H_1 \neq \bar{H}_1$, we see immediately that $x < 1$ and $y = 1$. This is not possible. Therefore, (14a) must hold.

Define H_1 and H_2 as in (13a) and (13b), respectively, with $\bar{k}_2 = 0$ and $\bar{k}_1$ satisfying (10d). Then such H_1 and H_2 are the positive solutions of (1) satisfying (7a). Now, if we can show that the positive solutions of equation (1) satisfying (7a) are unique, then the proof of the theorem will be complete. To this end, we note, as observed in the first paragraph of the proof, that $x_{\min}$ must be equal to 1. Therefore, $\int_{-\alpha}^{1}(H_1(\mu) - H_{1,\min}(\mu))\,d\mu = 0$, and so $H_1 \equiv H_{1,\min}$ and $H_2 \equiv H_{2,\min}$. Thus, the theorem is proved.

We conclude this paper with the following remarks.

Remarks. 1. We may conclude, via the proofs of Theorems 3 and 4, that for $c = 1$, if x and y are not both equal to 1, then equation (1) admits exactly two positive solutions.

2. On the other hand, if $x = y = 1$, then equation (1) has unique positive solutions. To see this, we note that either $x_{\min} := \frac{1}{2}\int_{\alpha}^{1} H_{1,\min}(\mu)\,d\mu$ or $y_{\min} := \frac{1}{2}\int_{-\alpha}^{1} H_{2,\min}(\mu')\,d\mu'$ is equal to 1. We assume, without loss of generality, that $y_{\min} = 1$. Thus,

$$0 = \frac{1}{2}\int_{-\alpha}^{1} H_{2,\min}(\mu') - H_2(\mu')\,d\mu'.$$

Since $H_{2,\min} - H_2$ is a continuous nonpositive function, we find that $H_{2,\min} \equiv H_2$, and hence $H_{1,\min} \equiv H_1$. Note that for $\alpha = 0$ and $c = 1$, we have $x = y = 1$.

3. The case where c is not a constant can be easily generalized.

Acknowledgments. The author would like to thank the referees for their helpful comments.

REFERENCES

[1] I. W. BUSBRIDGE, *On the H-function of Chandrasekhar*, Quart. J. Math. Oxford Ser., 8 (1957), pp. 133–140.

[2] ———, *On solutions of Chandrasekhar's integral equation*, Trans. Amer. Math. Soc., 105 (1962), pp. 112–117.

[3] S. CHANDRASEKHAR, *Radiative Transfer*, Dover, New York, 1960.

[4] ———, *The transfer of radiation in stellar atmospheres*, Bull. Amer. Math. Soc., 53 (1947), pp. 641–711.

[5] F. CORON, *Computation of the asymptotic states for linear half space kinetic problem*, Transport Theory Statist. Phys., 19 (1990), pp. 89–114.
[6] M. M. CRUM, *On an integral equation of Chandrasekhar*, Quart. J. Math. Oxford Ser., 18 (1947), pp. 244–252.
[7] C. FOX, *A solution of Chandrasekhar's integral equation*, Trans. Amer. Math. Soc., 99 (1961), pp. 285–291.
[8] B. D. GANAPOL, *An investigation of a simple transport model*, Transport Theory Statist. Phys., 21 (1992), pp. 1–37.
[9] V. HUSTON AND J. S. PYM, *Applications of Functional Analysis and Operator Theory*, Academic Press, New York, 1980.
[10] R. W. LEGGETT, *A new approach to the H-equation of Chandrasekhar*, SIAM J. Math. Anal., 7 (1976), pp. 542–550.
[11] C. A. STUART, *Existence theorems for a class of nonlinear integral equaitons*, Math. Z., 137 (1974), pp. 49–66.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 880–924, July 1995

006

ANALYSE SPECTRALE D'UNE BANDE ACOUSTIQUE MULTISTRATIFIÉE I: PRINCIPE D'ABSORPTION LIMITE POUR UNE STRATIFICATION SIMPLE*

ELISABETH CROC† ET YVES DERMENJIAN†

Abstract. One considers the operator $A = -\nabla.c^2\nabla$ governing the wave propagation in an acoustic strip $\Omega = \{(x,z) \in I\!R^2 \ / \ 0 < z < H\}$ with Neumann condition at $z = 0$ and Dirichlet condition at $z = H$. The celerity c describes the stratification of the medium: it is a measurable, piecewise constant function, with a finite number of strictly positive values. In this first paper, the spectral analysis is developed for the so-called free operators associated with a simple stratification: the celerity depends in a first case only on the variable x and in a second case on the variable z. A complete set of generalized eigenfunctions for the operator A is explicited. A limiting absorption principle is then deduced for each point of the spectrum, even at the bottom of the essential spectrum.

Key words. stratified medium, acoustic waves, self-adjoint operator, spectral analysis, threshold, limiting absorption principle

AMS subject classifications. 35L05, 35P, 47A70

Résumé. On s'intéresse à l'opérateur $A = -\nabla.c^2\nabla$ régissant la propagation des ondes acoustiques dans une bande $\Omega = \{(x,z) \in I\!R^2 \ / \ 0 < z < H\}$, avec conditions limites de Neumann en $z = 0$ et de Dirichlet en $z = H$. La vitesse c rend compte de la stratification du milieu : c'est une fonction mesurable, constante par pavés, prenant un nombre fini de valeurs strictement positives. Dans ce premier article, on fait l'analyse spectrale des opérateurs "libres" correspondant à un milieu stratifié dans une seule direction : la vitesse ne dépend donc que d'une variable, x dans un premier cas et z dans un deuxième cas. On construit explicitement un système complet de fonctions propres généralisées pour l'opérateur A. On en déduit un principe d'absorption limite en tout point du spectre, sans exclure la borne inférieure du spectre essentiel.

Mots clés. milieu stratifié, ondes acoustiques, opérateur auto-adjoint, analyse spectrale, seuil, principe d'absorption limite

1. Introduction.

Le principe d'absorption limite est largement utilisé dans les méthodes stationnaires. Ces dernières ont montré leur efficacité dans l'étude de la matrice de diffusion S de nombreux opérateurs auto-adjoints si on les considère comme des opérateurs perturbés d'opérateurs plus simples dits libres. La littérature et les exemples sont nombreux lorsque l'on combine un principe d'absorption limite soit avec des conditions de radiation (Lyford [L], Wilcox [Wi75], Eidus [E69], etc.), soit avec un théorème de division, que ce soit pour des perturbations à courte portée (Agmon [A], Dermenjian et Guillot [DG86], Hörmander [Hö], Weder [We], etc.) ou à longue portée.

Lorsque l'opérateur libre est suffisamment simple, une première étape dans l'obtention d'un théorème d'absorption limite est la description des propriétés d'un système complet de fonctions propres généralisées si on ne connaît pas de fonction de Green. Cette approche spectrale conditionne la suite et l'on comprend que, arrivés à ce point, de nombreux auteurs (Guillot [G], Guillot et Wilcox [GW], etc.) considèrent avoir fait l'essentiel. Remarquons que la notion d'opérateur simple est essentiellement subjective puisque des principes d'absorption limite et d'amplitude limite ont été obtenus dans des situations très diverses: Ben Artzi et Devinatz [BD], Ben Artzi, Dermenjian et

* Received by the editors May 5, 1993; accepted for publication (in revised form) December 8, 1993. Ce travail a été subventionné partiellement par Elf Aquitaine Production.

† UFR MIM de l' Université de Provence, Centre de Mathématiques et d'Informatique, Technopôle de Château-Gombert, 39 rue Joliot-Curie, F 13453 Marseille Cedex 13, France (ecroc@gyptis.univ-mrs.fr et dermenji@gyptis.univ-mrs.fr).

Guillot [BDG], DeBièvre et Pravica [DP], Dermenjian et Guillot [DG88], Eidus [E86], Hachem [Ha], Kikuchi et Tamura [KT], Tamura [T], etc.

Ce premier article traite deux situations simples mais qui semblent avoir été peu étudiées en mathématiques, si on met à part Morgenröther et Werner ([W87], [MW87], etc.) et quelques autres. Ces situations faciliteront la compréhension de la démarche suivie qui ressemble à celle utilisée par Wilcox [Wi84] lorsqu'il construit un système complet de fonctions propres généralisées. Nous nous en sommes écartés ensuite pour établir le principe d'absorption limite. Ces deux études permettront l'examen ultérieur d'un modèle plus réaliste, avec deux stratifications accolées. Ce dernier modèle, multistratifié, sera l'objet d'un prochain article.

Un des objectifs visés par cet article est aussi de mettre en place les outils permettant d'obtenir, à l'aide d'un principe d'absorption limite, une fonction de Green qui soit utilisable pour les applications numériques. On peut se rapporter au rapport [CD] pour cette question qui sera développée plus tard avec des résultats numériques.

La modélisation d'un problème particulier de sismique conduit à étudier l'équation d'ondes scalaire

$$\partial_t^2 v - \nabla.(c^2(x,z)\nabla v) = S \tag{1.1}$$

dans la bande infinie

$$\Omega = \{(x,z) \in I\!R^2 \ / \ z \in (0,H)\}. \tag{1.2}$$

L'opérateur ∇ désigne l'opérateur gradient $(\partial_x, \partial_z)^t$. La fonction v représente le déplacement. La source S est donnée. La fonction c, mesurable, est un profil de vitesse qui rend compte de la stratification du milieu, et elle admet un minorant c_m strictement positif et un majorant c_M. Les conditions limites (CL) sont fixées en $z = 0$ et $z = H$, conditions de Dirichlet ou de Neumann.

L'approche stationnaire associe aux équations (1.1) et (CL) l'opérateur

$$D(A) = \{u \in H^1(\Omega) \ / \ -\nabla.(c^2\nabla u) \in L^2(\Omega) \text{ et } u \text{ vérifie (CL)}\}, \tag{1.3a}$$

$$Au = -\nabla.(c^2\nabla u) \ \text{ si } \ u \in D(A). \tag{1.3b}$$

L'étude des solutions d'énergie finie de (1.1) passe par l'étude spectrale de l'opérateur $(D(A), A)$, auto-adjoint dans l'espace de Hilbert $L^2(\Omega)$.

Pour mener les calculs, nous avons choisi les conditions

$$c^2(x,0)\,\partial_z v(x,0,t) = 0 \quad \text{et} \quad v(x,H,t) = 0. \tag{1.4}$$

L'étude serait similaire avec des conditions limites de Dirichlet (respectivement Neumann), ce qui est le cas considéré dans [W87]. Les géophysiciens de la Société Elf Aquitaine qui nous ont posé ce problème, avaient déjà écrit un code de calcul pour une bande élastique, avec la condition de surface libre en $z = 0$, et la condition de fonds rigide en $z = H$, dont l'analogue en acoustique est (1.4). Ceci explique notre choix, qui permettra de comparer plus tard nos résultats théoriques avec la solution numérique obtenue par une autre approche.

Dans cet article, nous considérons des milieux simplement stratifiés ou encore stratifiés dans une seule direction. Les interfaces correspondant aux discontinuités du profil c sont donc parallèles aux axes Ox ou Oz.

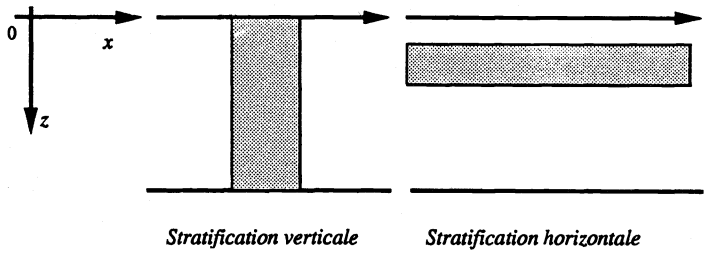


FIG. 1.1.

Dans §2, l'interface est située en $x = 0$ et la fonction c prend deux valeurs strictement positives et distinctes:

$$c(x,z) = c(x) = \begin{cases} c_1 \text{ si } x < 0, \\ c_2 \text{ si } x > 0. \end{cases} \tag{1.5}$$

Dans §3, l'interface est située en $z = h$ $(0 < h < H)$ et la fonction c prend deux valeurs strictement positives qui peuvent être éventuellement égales:

$$c(x,z) = c(z) = \begin{cases} c_1 \text{ si } z \in (0,h), \\ c_2 \text{ si } z \in (h,H). \end{cases} \tag{1.6}$$

La généralisation à des fonctions constantes par intervalles, qui prennent un nombre fini de valeurs strictement positives (cf. Fig. 1.1), ne présente pas de difficultés théoriques supplémentaires. Nous travaillons de fait dans le cadre plus général d'une fonction $c(z)$ satisfaisant

$$\textbf{(H)} \qquad c \in L^\infty((0,H)) \quad \text{et} \quad \operatorname{Min} c(z) \geq c_m > 0.$$

Un développement en fonctions propres généralisées, déterminé par séparation des variables, est utilisé pour démontrer un principe d'absorption limite pour A, et pour résoudre les problèmes aux limites associés à l'opérateur différentiel $(A - \mu I)$ lorsque μ décrit le plan complexe. On obtient ainsi les théorèmes 2.8 et 3.4. Remarquons que les théorèmes généraux (cf. [DL]) nous garantissent l'existence des fonctions propres généralisées, mais leur construction nécessite la connaissance de la mesure spectrale. Ici, nous pouvons déduire la mesure spectrale de la connaissance des fonctions propres généralisées.

A la lecture des deux principales conclusions de ce travail, données par les théorèmes 1.1 et 1.2, on notera un résultat qui n'est pas usuel chez les auteurs étudiant le Laplacien et ses perturbations : le principe d'absorption limite est obtenu en tout point de l'axe réel alors que le minimum du spectre essentiel est en général exclu. Ceci pose de manière naturelle la question du prolongement méromorphe à $\mathbb{C}$ de l'application $\lambda \mapsto (A - \lambda^2 I)^{-1}$ (cf. le théorème 1.1 de [SZ]).

Dans les démonstrations utilisées, comme beaucoup d'auteurs, nous avons introduit (cf. §§2.2.2 et 3.2), d'une part pour s réel, l'espace $L^2_s(\Omega)$ des fonctions f mesurables sur Ω telles que $(1 + x^2)^{s/2} f(x,z)$ soit de carré intégrable sur Ω, d'autre

part pour n entier positif et λ réel dans un intervalle I_n de la forme $[\lambda_n, +\infty)$, des opérateurs de trace $\tau_n(\lambda)$.

Pour la stratification verticale, c'est-à-dire pour le profil (1.5) avec $c_1 \neq c_2$, la nullité de ces opérateurs de trace aux seuils λ_n permet d'y établir un principe d'absorption limite avec des résolvantes $R^{\pm}(\lambda_n)$ définies sur l'espace $L^2_s(\Omega)$, $s > 1$, ce qui est habituel. Le principe obtenu (cf. le théorème 2.8) s'énonce comme suit.

THÉORÈME 1.1 (Stratification verticale). *Soit μ réel, appartenant au spectre de l'opérateur A défini par* (1.3) *avec le profil* (1.5). *Soit s réel vérifiant $s > 1$ si μ est un seuil λ_n de A et $s > 1/2$ sinon. On considère la résolvante $\zeta \mapsto R_A(\zeta) = (A - \zeta I)^{-1}$ comme une fonction définie sur $\mathbb{C}\backslash\sigma(A)$ à valeurs dans $B(L^2_s(\Omega), L^2_{-s}(\Omega))$.*

Alors les limites suivantes existent dans $B(L^2_s(\Omega), L^2_{-s}(\Omega))$ pour la topologie de la norme:

$$R^{\pm}_A(\mu) = \lim_{\zeta \to \mu,\ \pm \mathrm{Im}\zeta > 0} R_A(\zeta). \tag{1.7}$$

Il est à noter que la preuve du théorème 1.1 n'est pas valable lorsque $c_1 = c_2$. En effet, l'absence du mode propre (2.42), qui est un mode réfléchi sur l'interface $x = 0$, fait "exploser" la résolvante au voisinage des seuils λ_n (cf. la remarque 2.5 et le théorème 3.3).

Par contre pour la stratification horizontale, c'est-à-dire pour le profil (1.6), et plus généralement pour un profil satisfaisant l'hypothèse **(H)**, les opérateurs $\tau_n(\lambda)$ ne s'annulent plus aux seuils λ_n. Les résolvantes $R^{\pm}(\lambda_n)$ ne peuvent y être définies que pour $s > 1$, sur un sous-espace fermé de $L^2_s(\Omega)$, qui est toutefois dense dans $L^2(\Omega)$. Ce sous-espace de $L^2_s(\Omega)$ est un hyperplan, noté $NL^2_s(n)$, qui est introduit à la proposition 3.3. Le résultat est optimal (cf. le théorème 3.3 et la limite (3.55)). Nous l'utiliserons dans des applications à venir. Le principe d'absorption limite (cf. le théorème 3.4) est démontré pour des conditions limites (CL) de Dirichlet ou de Neumann en $z = 0$ ou $z = H$, et pour tout profil qui satisfait **(H)**. Il s'énonce comme suit.

THÉORÈME 1.2 (Stratification horizontale). *Soit μ réel, appartenant au spectre de l'opérateur A défini par* (1.3) *avec un profil $c(z)$ satisfaisant l'hypothèse* **(H)**. *Soit s réel, vérifiant $s > 1$ si μ est un seuil λ_n de A, et $s > 1/2$ sinon. Soit $E_s(\mu)$ l'espace de fonctions, égal à $NL^2_s(n)$ si μ est un seuil λ_n de A, et égal à $L^2_s(\Omega)$ sinon. Soit $E_s(\mu)'$ son dual topologique. On considère la résolvante $\zeta \mapsto R_A(\zeta) = (A - \zeta I)^{-1}$ comme une fonction définie sur $\mathbb{C}\backslash\sigma(A)$ à valeurs dans $B(E_s(\mu), E_s(\mu)')$.*

Alors les limites suivantes existent dans $B(E_s(\mu), E_s(\mu)')$ pour la topologie de la norme:

$$R^{\pm}_A(\mu) = \lim_{\zeta \to \mu,\ \pm \mathrm{Im}\zeta > 0} R_A(\zeta). \tag{1.8}$$

En particulier, le profil (1.6) avec $c_1 = c_2$ relève du théorème 1.2. On retrouve ainsi un résultat de Werner dans [W87]. Celui-ci considère le Laplacien dans un tube $\Omega = I\!R \times \Omega'$, avec des conditions limites de Dirichlet ou de Neumann sur $\partial\Omega$. Il montre que le principe d'absorption limite est en défaut (cf. (2.32) dans [W87]) au voisinage des valeurs propres λ_n de la section transversale Ω' du tube. Dans notre cas, ces valeurs propres sont celles de l'opérateur $(D(B_0), B_0)$ considéré à la proposition 3.1. Elles sont simples et associées aux fonctions propres $V(0, n, .)$. La formule (1.7) de [W87], qui explicite la réponse $v(x, z, t)$ à une source $S(x, z, t) = f(x, z)\ exp(-i\sqrt{\lambda_n}t)$, présente un terme résonant proportionnel à

$$\sqrt{t}\ \exp(-i\sqrt{\lambda_n}\, t)\ V(0, n, z) \int_\Omega f(x, z')\ V(0, n, z')\ dxdz'.$$

Celui-ci est nul si et seulement si

$$\int_\Omega f(x,z)\, V(0,n,z)\, dxdz = 0, \tag{1.9}$$

c'est-à-dire, selon (3.34), si et seulement si f est dans l'hyperplan $NL^2_s(n) = E_s(\lambda_n)$.

Une autre méthode pour traiter (1.1) et (1.4) consiste à travailler à fréquence fixée. C'est l'approche utilisée par les géophysiciens avec qui nous travaillons, en particulier Boelle dans [B]. Nous l'avons exposée pour le cas d'une stratification horizontale dans [CD]. Elle est par ailleurs bien adaptée à l'étude numérique des fonctions de Green, qui sont présentées dans ce rapport [CD].

2. Profil de vitesse c(x,z)=c(x) prenant deux valeurs distinctes. Le profil de vitesse, dans cette section, est le profil (1.5), c'est-à-dire indépendant de la variable z et prenant deux valeurs, distinctes et strictement positives, c_1 ou c_2 selon que la variable x est négative ou positive. On peut toujours supposer, quitte à changer l'orientation de l'axe $x'Ox$, que c_1 est supérieur à c_2.

Nous développons ci-après la théorie spectrale de l'opérateur associé au problème, puis nous établissons un principe d'absorption limite.

2.1. Théorie spectrale de l'opérateur $A = -\nabla \cdot (c^2(x)\nabla)$. Aux conditions limites (1.4) et à l'équation des ondes (1.1) sur l'ouvert Ω, nous associons la forme sesquilinéaire

$$a(u,v) = \int_\Omega c^2(x)\, (\partial_x u \partial_x \overline{v} + \partial_z u \partial_z \overline{v})\ dx\, dz.$$

Soit $V = \{v \in H^1(\Omega)\ /\ v|_{z=H} = 0\}$, muni de la norme induite par celle de $H^1(\Omega)$, pour laquelle V est un espace de Hilbert.

La forme a est définie et continue sur $V \times V$, et V-coercive par rapport à $L^2(\Omega)$. Lorsque le profil c est, par exemple, continu par morceaux, la théorie variationnelle lui associe l'opérateur non borné

$$A = -\partial_x(c^2(x)\partial_x) - c^2(x)\partial_z^2 \tag{2.1}$$

de domaine

$$D(A) = \{u \in H^1(\Omega)\ /\ Au \in L^2(\Omega)\ \text{ et }\ (c^2\partial_z u)\,|_{z=0} = u|_{z=H} = 0\} \tag{2.2}$$

qui est auto-adjoint positif dans $L^2(\Omega)$.

2.1.1. Séparation des variables. Les notations $(\cdot/\cdot)$ et $\|\cdot\|$ désignent respectivement le produit scalaire et la norme dans l'espace de Hilbert approprié. Rappelons le théorème de diagonalisation d'un opérateur A auto-adjoint sur un espace de Hilbert $\mathcal{H}$ séparable qui permet d'expliciter les résolvantes et la famille spectrale de l'opérateur. Nous renvoyons à Reed et Simon [RS] ou Dautray et Lions [DL] pour plus de détails.

THÉORÈME 2.1. *Etant donné un opérateur A auto-adjoint sur un espace de Hilbert $\mathcal{H}$ séparable, il existe un espace de Hilbert $\widetilde{\mathcal{H}}$, intégrale hilbertienne sur $\mathbb{R}$ d'espaces de Hilbert $H(\lambda)$ pour une mesure $d\mu$ sur $\mathbb{R}$, soit*

$$\widetilde{u} = (\widetilde{u}(\lambda))_{\lambda\in\mathbb{R}} \in \widetilde{\mathcal{H}} \Longleftrightarrow \int_{\mathbb{R}} \|\widetilde{u}(\lambda)\|^2_{H(\lambda)} d\mu(\lambda) < +\infty \tag{2.3a}$$

avec

$$\|\widetilde{u}\|^2_{\widetilde{\mathcal{H}}} = \int_{\mathbb{R}} \|\widetilde{u}(\lambda)\|^2_{H(\lambda)} d\mu(\lambda), \tag{2.3b}$$

et une transformation unitaire $\widetilde{\mathcal{F}}$ de $\mathcal{H}$ sur $\widetilde{\mathcal{H}}$ qui réduit l'opérateur A, soit

$$u \in D(A) \Longleftrightarrow \int_{\mathbb{R}} \lambda^2 \|\widetilde{\mathcal{F}}u(\lambda)\|^2 d\mu(\lambda) < +\infty \tag{2.4a}$$

avec

$$\widetilde{\mathcal{F}}(Au)(\lambda) = \lambda \widetilde{\mathcal{F}}u(\lambda). \tag{2.4b}$$

La famille spectrale $(\Pi_A(\lambda))_{\lambda \in \mathbb{R}}$ de l'opérateur A est alors déterminée par

$$\forall f \in \mathcal{H}\ ,\ (\Pi_A(\lambda)f/f)_{\mathcal{H}} = \int_{-\infty}^{\lambda} \|\widetilde{\mathcal{F}}f(t)\|^2 d\mu(t). \tag{2.5}$$

En particulier on peut décrire le spectre $\sigma(A)$ et préciser sa nature grâce à la connaissance de $\widetilde{\mathcal{F}}$. $\widetilde{\mathcal{F}}$ est appelée une représentation spectrale de A sur l'espace $\widetilde{\mathcal{H}}$.

COROLLAIRE 2.1. *Pour $\zeta \notin \sigma(A)$ et $f \in \mathcal{H}$, on a*

$$R_A(\zeta)f = (A - \zeta I)^{-1} f = \widetilde{\mathcal{F}}^{-1}\left(\frac{\widetilde{\mathcal{F}}f(\lambda)}{\lambda - \zeta}\right). \tag{2.6}$$

La séparation des variables x et z, dans l'ouvert $\Omega = \mathbb{R} \times (0, H)$ et dans l'espace $\mathcal{H} = L^2(\Omega; dxdz)$ permet de ramener l'analyse spectrale de l'opérateur A défini par (2.1) et (2.2), à coefficients indépendants de z, à celle d'une suite d'opérateurs B_n auto-adjoints dans $L^2(\mathbb{R}; dx)$. Plus précisément, l'espace de Hilbert $L^2(\Omega; dxdz)$ s'identifie au produit tensoriel hilbertien $L^2(\mathbb{R}; dx) \otimes L^2((0, H); dz)$. Cette identification peut être explicitée avec une base de $L^2((0, H); dz)$.

PROPOSITION 2.1. *Soit $(V_n)_{n \geq 1}$ une base orthonormée de $L^2((0, H))$. Une fonction u appartient à l'espace $L^2(\Omega)$ si et seulement si elle peut s'écrire*

$$u(x, z) = \sum_{n \geq 1} u_n(x) V_n(z), \text{ série convergente dans l'espace } L^2(\Omega; dxdz), \tag{2.7a}$$

avec

$$u_n(x) = (u(x, .)/V_n)_{L^2((0,H);dz)} \text{ et } \|u\|^2_{L^2(\Omega;dxdz)} = \sum_{n \geq 1} \|u_n\|^2_{L^2(\mathbb{R};dx)}. \tag{2.7b}$$

Choisissons pour base de $L^2((0, H))$, une base orthonormée de fonctions propres pour l'opérateur $(D(A_2), A_2)$ défini par

$$D(A_2) = \{v \in H^1((0, H)) \,/\, v'' \in L^2((0, H)) \quad \text{et} \quad v'(0) = v(H) = 0\}, \tag{2.8a}$$

$$A_2 u = -u'' \text{ si } u \in D(A_2). \tag{2.8b}$$

Cet opérateur, auto-adjoint dans $L^2((0, H))$, étant à inverse compact, une telle base existe. Le calcul donne

$$n \in \mathbb{N}^* \quad \text{et} \quad -V_n'' = q_n^2 V_n, \tag{2.9a}$$

$$q_n = \frac{1}{H}[\frac{\pi}{2} + (n-1)\pi] \quad \text{et} \quad V_n(z) = \sqrt{\frac{2}{H}} \sin [q_n(H-z)]. \tag{2.9b}$$

PROPOSITION 2.2. *Avec les notations précédentes, si on choisit la base* (2.9) *de* $L^2((0,H))$ *dans la décomposition* (2.7) *de* $L^2(\Omega)$, *l'opérateur défini par* (2.1) *et* (2.2) *est la somme directe des opérateurs suivants, auto-adjoints dans* $L^2(\mathbb{R})$,

$$n \in \mathbb{N}^* \quad et \quad D(B_n) = \{u \in H^1(\mathbb{R}) \,/\, B_n u \in L^2(\mathbb{R})\}, \tag{2.10a}$$

$$B_n u = -\left(c^2 u'\right)' + c^2 q_n^2 u \quad si \quad u \in D(B_n), \tag{2.10b}$$

soit encore pour toute fonction u *de* $D(A)$,

$$u(x,z) = \sum_{n\geq 1} u_n(x) V_n(z) \quad et \quad Au(x,z) = \sum_{n\geq 1} (B_n u_n)(x) V_n(z). \tag{2.11}$$

THÉORÈME 2.2. *Avec les notations précédentes, soit* n *entier positif et*

$$\widetilde{\mathcal{F}}_n \begin{cases} L^2(\mathbb{R}; dx) \to \widetilde{\mathcal{H}}_n \\ v \longmapsto \widetilde{\mathcal{F}}_n v \end{cases}$$

une représentation spectrale de l'opérateur B_n. *Alors on définit une représentation spectrale de l'opérateur* A *sur la somme directe hilbertienne*

$$\widetilde{\mathcal{H}} = \bigoplus_{n\geq 1} \widetilde{\mathcal{H}}_n$$

par

$$\widetilde{\mathcal{F}} \begin{cases} L^2(\Omega) \to \widetilde{\mathcal{H}} \\ u \longmapsto \widetilde{\mathcal{F}} u \end{cases} \tag{2.12a}$$

$$\widetilde{\mathcal{F}} u = (\widetilde{\mathcal{F}}_n u_n)_{n\geq 1} \quad et \quad u_n(x) = (u(x,.)/V_n)_{L^2((0,H))}. \tag{2.12b}$$

Nous explicitons au théorème 4.4 de l'annexe une représentation spectrale $\widetilde{\mathcal{F}}_n$ pour l'opérateur B_n dans le cas $c_1 > c_2$.

COROLLAIRE 2.2. *Le spectre* $\sigma(A)$ *est donné par*

$$\sigma(A) = \overline{\bigcup_{n\geq 1} \sigma(B_n)}. \tag{2.13}$$

Pour toutes fonctions f *et* g *de* $L^2(\Omega)$, *et pour tout nombre complexe* $\zeta \notin \sigma(A)$, *la résolvante* $R_A(\zeta) = (A - \zeta I)^{-1}$ *vérifie*

$$(R_A(\zeta) f/g)_{L^2(\Omega)} = \sum_{n\geq 1} \int_{\mathbb{R}} \frac{\widetilde{f}(\lambda,n)\overline{\widetilde{g}(\lambda,n)}}{\lambda - \zeta} d\mu_n(\lambda) \tag{2.14}$$

avec $\widetilde{f}(\lambda,n) = (\,\widetilde{\mathcal{F}}_n\,(f(x,.)/V_n)_{L^2((0,H))}\,)(\lambda)$.

2.1.2. Description explicite d'une représentation spectrale. On rappelle l'hypothèse $c_1 > c_2$, et la définition des fonctions propres généralisées de A : ce sont des fonctions $\psi = \psi(x,z)$ vérifiant

$$\text{(2.15)} \qquad \begin{cases} A\psi = \lambda\psi \text{ dans } \mathcal{D}'(\Omega), \text{ avec } \lambda \text{ convenable dans } \mathbb{R}^+, \\ \psi \text{ est localement dans } D(A), \\ \psi \text{ est bornée dans } \Omega. \end{cases}$$

On déduit du théorème 2.2 et des formules (4.2), (4.4) et (4.12) établies dans l'annexe, l'analyse qui suit pour l'opérateur A.

THÉORÈME 2.3 (Fonctions propres généralisées pour A). *Ayant défini en* (2.9) *la suite des éléments propres* q_n^2 *et* $V_n(z)$, $n \geq 1$, *on note*

$$\text{(2.16a)} \qquad \xi_1 = \xi_1(\lambda, n) = \frac{1}{c_1}(\lambda - c_1^2 q_n^2)^{1/2} \quad \textit{si } \lambda > c_1^2 q_n^2,$$

$$\text{(2.16b)} \qquad \xi_2 = \xi_2(\lambda, n) = \frac{1}{c_2}(\lambda - c_2^2 q_n^2)^{1/2} \quad \textit{si } \lambda > c_2^2 q_n^2,$$

$$\text{(2.16c)} \qquad \xi_1' = \xi_1'(\lambda, n) = \frac{1}{c_1}(c_1^2 q_n^2 - \lambda)^{1/2} \quad \textit{si } \lambda < c_1^2 q_n^2,$$

$$\text{(2.17)} \qquad I_n = (c_2^2 q_n^2, c_1^2 q_n^2) \quad \textit{et} \quad J_n = (c_1^2 q_n^2, +\infty).$$

Les fonctions $\psi^i = \psi^i(\lambda, n, x, z)$, $i = 0, 1$ *ou* 2, $n \geq 1$, *définies sur* Ω *par*

$$\text{(2.18a)} \qquad \psi^0(\lambda, n, x, z) = \varphi^0(\lambda, n, x) V_n(z) \quad \textit{si } \lambda \in I_n,$$

$$\text{(2.18b)} \qquad \psi^1(\lambda, n, x, z) = \varphi^1(\lambda, n, x) V_n(z) \quad \textit{si } \lambda \in J_n,$$

$$\text{(2.18c)} \qquad \psi^2(\lambda, n, x, z) = \varphi^2(\lambda, n, x) V_n(z) \quad \textit{si } \lambda \in J_n.$$

avec

$$\text{(2.19a)} \qquad \varphi^0(\lambda, n, x) = \left(\frac{c_2^2 \xi_2}{\pi[(c_1^2\xi_1')^2 + (c_2^2\xi_2)^2]}\right)^{1/2} \begin{cases} e^{\xi_1' x} & \textit{si } x < 0, \\ \cos(\xi_2 x) + \dfrac{c_1^2 \xi_1'}{c_2^2 \xi_2} \sin(\xi_2 x) & \textit{si } x > 0, \end{cases}$$

$$\text{(2.19b)} \qquad \varphi^1(\lambda, n, x) = \left(\frac{c_2^2 \xi_2}{\pi c_1^2 \xi_1 (c_1^2 \xi_1 + c_2^2 \xi_2)}\right)^{1/2} \begin{cases} \sin(\xi_1 x) & \textit{si } x < 0, \\ \dfrac{c_1^2 \xi_1}{c_2^2 \xi_2} \sin(\xi_2 x) & \textit{si } x > 0, \end{cases}$$

$$\text{(2.19c)} \qquad \varphi^2(\lambda, n, x) = \left(\frac{1}{\pi(c_1^2 \xi_1 + c_2^2 \xi_2)}\right)^{1/2} \begin{cases} \cos(\xi_1 x) & \textit{si } x < 0, \\ \cos(\xi_2 x) & \textit{si } x > 0, \end{cases}$$

sont des fonctions propres généralisées de A. *Elles sont réelles et vérifient l'équation différentielle* $A\psi = \lambda\psi$ *dans* $\mathcal{D}'(\Omega)$. *Elles forment une famille complète au sens où elles permettent de construire une représentation spectrale* $\widetilde{\mathcal{F}}$ *de* A.

THÉORÈME 2.4 (Représentation spectrale $\widetilde{\mathcal{F}}$ de A). *Soit f une fonction de $L^2(\Omega)$. Pour n entier strictement positif, pour $i = 0$ et presque tout $\lambda \in I_n$, ainsi que pour $i = 1$ ou 2 et presque tout $\lambda \in J_n$, on peut définir les coefficients de Fourier généralisés de f par*

$$\widetilde{f}^i(\lambda, n) = L^2(d\lambda) - \lim_{X \to +\infty} \int_{\Omega \cap \{(x,z)/|x|<X\}} f(x,z)\overline{\psi^i(\lambda, n, x, z)}dxdz. \tag{2.20}$$

Ils vérifient $\widetilde{f}^i(.,n) \in L^2(I_n)$ si $i = 0$, $\widetilde{f}^i(.,n) \in L^2(J_n)$ si $i = 1$ ou 2, et

$$\widetilde{f} = (\widetilde{f}^0, \widetilde{f}^1, \widetilde{f}^2) \in \widetilde{\mathcal{H}} = \bigoplus_{n \geq 1} \widetilde{\mathcal{H}}_n \quad \text{avec } \widetilde{\mathcal{H}}_n = L^2(I_n) \oplus L^2(J_n) \oplus L^2(J_n).$$

La transformation $[f \in L^2(\Omega) \mapsto \widetilde{f} \in \widetilde{\mathcal{H}}]$ est unitaire. L'égalité de Bessel-Parseval pour f et g fonctions de $L^2(\Omega)$ s'écrit

$$(f/g)_{L^2(\Omega)} = \sum_{n \geq 1} \left[\int_{I_n} \widetilde{f}^0(\lambda, n)\overline{\widetilde{g}^0(\lambda, n)}d\lambda + \sum_{i=1}^{2} \int_{J_n} \widetilde{f}^i(\lambda, n)\overline{\widetilde{g}^i(\lambda, n)}d\lambda \right].$$

Le développement en fonctions propres généralisées de f fonction de $L^2(\Omega)$ s'écrit

$$f(x,z) = \sum_{n \geq 1} \left[\int_{I_n} \widetilde{f}^0(\lambda, n)\psi^0(\lambda, n, x, z)d\lambda + \sum_{i=1}^{2} \int_{J_n} \widetilde{f}^i(\lambda, n)\psi^i(\lambda, n, x, z)d\lambda \right], \tag{2.21}$$

la convergence pour chaque intégrale en λ ayant lieu dans l'espace $L^2(\mathbb{R}; dx)$ et la sommation de la série se faisant au sens de la norme hilbertienne dans $L^2(\Omega; dxdz)$. Si f appartient à $D(A)$, on a la formule de diagonalisation

$$\widetilde{Af}^i(\lambda, n) = \lambda \widetilde{f}^i(\lambda, n). \tag{2.22}$$

COROLLAIRE 2.3. *Le spectre $\sigma(A)$ est absolument continu. Il est donné par*

$$\sigma(A) = \bigcup_{n \geq 1} [c_2^2 q_n^2, +\infty) = \left[\frac{c_2^2 \pi^2}{4H^2}, +\infty \right). \tag{2.23}$$

Pour toutes fonctions f et g de $L^2(\Omega)$, et pour tout nombre complexe $\zeta \notin \sigma(A)$, la résolvante $R_A(\zeta) = (A - \zeta I)^{-1}$ vérifie

$$(R_A(\zeta)f/g)_{L^2(\Omega)} = \sum_{n \geq 1} \left[\int_{I_n} \frac{\widetilde{f}^0(\lambda, n)\overline{\widetilde{g}^0(\lambda, n)}}{\lambda - \zeta} d\lambda + \sum_{i=1}^{2} \int_{J_n} \frac{\widetilde{f}^i(\lambda, n)\overline{\widetilde{g}^i(\lambda, n)}}{\lambda - \zeta} d\lambda \right]. \tag{2.24}$$

Remarque 2.1. La normalisation choisie nous permet de prendre la mesure de Lebesgue $d\lambda$ pour les mesures $d\mu$ ou $d\mu_n$.

2.1.3. Une deuxième transformation unitaire. Des changements de variables, mieux adaptés à la preuve d'un principe d'absorption limite pour A comme nous le verrons dans §2.2 ci-après, permettent de paramétriser la famille des fonctions

propres généralisées $\psi^i(\lambda, n, x, z)$, décrite au théorème 2.3, à l'aide d'une nouvelle variable ξ strictement positive.

Pour cela, nous posons

$$a = \frac{1}{c_2}(c_1^2 - c_2^2)^{1/2}, \tag{2.25}$$

$$\begin{aligned} \Psi^1(\xi, n, x, z) &= \phi^1(\xi, n, x) V_n(z) \\ &= c_1\sqrt{2\xi}\psi^1(c_1^2(\xi^2 + q_n^2), n, x, z) \quad \text{pour } \xi > 0, \end{aligned} \tag{2.26}$$

$$\begin{aligned} \Psi^2(\xi, n, x, z) &= \phi^2(\xi, n, x) V_n(z) \\ &= c_2\sqrt{2\xi}\begin{cases} \psi^0(c_2^2(\xi^2 + q_n^2), n, x, z) & \text{pour } \xi \in (0, aq_n], \\ \psi^2(c_2^2(\xi^2 + q_n^2), n, x, z) & \text{pour } \xi \geq aq_n. \end{cases} \end{aligned} \tag{2.27}$$

Pour $\xi = aq_n$, (2.27) définit correctement Ψ^2 ou ϕ^2. En effet les formules (2.19a) et (2.19c), écrites avec $\lambda = c_2^2(a^2q_n^2 + q_n^2) = c_1^2q_n^2$, valeur de λ pour laquelle $\xi_1(\lambda, n) = \xi_1'(\lambda, n) = 0$ et $\xi_2(\lambda, n) = aq_n$, définissent une seule fonction

$$\varphi(c_1^2q_n^2, n, x) = \varphi^0(c_1^2q_n^2, n, x) = \varphi^2(c_1^2q_n^2, n, x) = \left(\frac{1}{\pi c_2^2 aq_n}\right)^{1/2}\begin{cases} 1 \text{ si } x < 0, \\ \cos(aq_n x) \text{ si } x > 0. \end{cases}$$

Cette fonction φ engendre l'espace propre généralisé associé à $\lambda = c_1^2q_n^2$ de l'opérateur B_n (cf. le théorème 4.1 de l'annexe et la formule (4.8)). Les formules (2.18), écrites avec $\lambda = c_1^2q_n^2$, définissent donc une seule fonction propre généralisée de l'opérateur A, égale à

$$\psi(c_1^2q_n^2, n, x, z) = \psi^0(c_1^2q_n^2, n, x, z) = \psi^2(c_1^2q_n^2, n, x, z) = \varphi(c_1^2q_n^2, n, x)V_n(z).$$

La définition (2.27) de Ψ^2 ou ϕ^2 est ainsi validée pour $\xi = aq_n$ et on a

$$\Psi^2(aq_n, n, x, z) = \phi^2(aq_n, n, x)V_n(z) = \sqrt{\frac{2}{\pi}}V_n(z)\begin{cases} 1 \text{ si } x < 0, \\ \cos(aq_n x) \text{ si } x > 0. \end{cases} \tag{2.28}$$

Pour $n \geq 1, i = 1$ ou 2 et $\xi > 0$, nous avons

$$A\Psi^i(\xi, n, ., .) = \lambda_i(\xi, n)\Psi^i(\xi, n, ., .) \quad \text{avec } \lambda_i(\xi, n) = c_i^2(\xi^2 + q_n^2), \tag{2.29}$$

et nous pouvons énoncer des résultats parallèles à ceux du théorème 2.4.

THÉORÈME 2.5 (Deuxième représentation spectrale de A). *Soit f une fonction de $L^2(\Omega)$. Pour n entier strictement positif, pour $i = 1$ ou 2 et pour presque tout réel $\xi > 0$, on définit les coefficients*

$$f^i(\xi, n) = L^2((0, +\infty); d\xi) - \lim_{X \to +\infty} \int_{\Omega \cap \{(x,z)/|x|<X\}} f(x, z)\overline{\Psi^i(\xi, n, x, z)}dxdz. \tag{2.30}$$

Ces coefficients sont associés aux fonctions propres généralisées $\Psi^i(\xi, n, ., .)$, définies en (2.26) et (2.27) à partir des fonctions $\psi^i(\lambda, n, ., .)$, et réelles. Ils vérifient $f^i(., n) \in L^2((0, +\infty))$ et

$$(f^1, f^2) \in \bigoplus_{n \geq 1} \mathcal{H}_n \text{ avec } \mathcal{H}_n = L^2((0, +\infty)) \oplus L^2((0, +\infty)). \tag{2.31}$$

La transformation $[f \in L^2(\Omega) \longmapsto (f^1, f^2) \in \bigoplus_{n\geq 1} \mathcal{H}_n]$ *est unitaire et on a*

$$f(x,z) = \sum_{n\geq 1} \sum_{i=1}^{2} \int_0^{+\infty} f^i(\xi,n)\Psi^i(\xi,n,x,z)d\xi, \tag{2.32a}$$

$$(f/g)_{\mathcal{H}} = \sum_{n\geq 1} \sum_{i=1}^{2} \int_0^{+\infty} f^i(\xi,n)\overline{g^i(\xi,n)}d\xi \quad \textit{pour } f,g \in L^2(\Omega). \tag{2.32b}$$

Si f *appartient à* $D(A)$, *on a*

$$(Af)^i(\xi,n) = \lambda_i(\xi,n)f^i(\xi,n), \quad \textit{avec } \lambda_i(\xi,n) \textit{ défini en } (2.31). \tag{2.33}$$

Pour toutes fonctions f *et* g *de* $L^2(\Omega)$, *et pour tout nombre complexe* $\zeta \notin \sigma(A)$, *la résolvante* $R_A(\zeta) = (A - \zeta I)^{-1}$ *vérifie*

$$(R_A(\zeta)f/g)_{L^2(\Omega)} = \sum_{n\geq 1} \sum_{i=1}^{2} \int_0^{+\infty} \frac{f^i(\xi,n)\overline{g^i(\xi,n)}}{c_i^2(\xi^2+q_n^2)-\zeta}d\xi. \tag{2.34}$$

Remarque 2.2. Cette paramétrisation en ξ des fonctions propres généralisées

$$\Psi^i(\xi,n,x,z) = \phi^i(\xi,n,x)V_n(z), \qquad i = 1 \text{ ou } 2,$$

met clairement en évidence la dissymétrie du milieu de propagation: le mode Ψ^1 se propage aussi bien dans la demi-bande $\Omega^- = \{(x,z) \in \Omega / x < 0\}$ que dans la demi-bande $\Omega^+ = \{(x,z) \in \Omega / x > 0\}$, alors que le mode Ψ^2 n'a ce comportement que si $\xi > aq_n$. Pour $\xi < aq_n$, c'est un mode qui s'amortit rapidement dans Ω^- selon la direction normale à l'interface $x = 0$ (on pourra aussi se reporter aux expressions (2.49) pour $\phi^1(\xi,n,x)$, (2.42) et(2.43) pour $\phi^2(\xi,n,x)$).

Remarque 2.3. La représentation spectrale pour une bande homogène, où $c_1 = c_2 = c$, est aussi obtenue par ces calculs. L'analyse du théorème 2.3 se fait à l'aide de deux familles de fonctions, indexées par $n \geq 1$ et $\lambda > cq_n^2$, données pour x réel et pour $\xi = \frac{1}{c}(\lambda - c^2q_n^2)^{1/2}$, par

$$\varphi^1(\lambda,n,x) = \left(\frac{1}{2\pi c^2\xi}\right)^{1/2} \sin(\xi x), \tag{2.35}$$

$$\varphi^2(\lambda,n,x) = \left(\frac{1}{2\pi c^2\xi}\right)^{1/2} \cos(\xi x). \tag{2.36}$$

La transformation du théorème 2.5 est alors associée aux fonctions propres généralisées, indexées par $\xi > 0$ et $n \geq 1$, et définies pour $(x,z) \in \Omega$, par

$$\Psi^1(\xi,n,x,z) = \phi^1(\xi,n,x)\ V_n(z) = \frac{1}{\sqrt{\pi}} \sin(\xi x)\ V_n(z), \tag{2.37}$$

$$\Psi^2(\xi,n,x,z) = \phi^2(\xi,n,x)\ V_n(z) = \frac{1}{\sqrt{\pi}} \cos(\xi x)\ V_n(z). \tag{2.38}$$

Elle est naturellement attachée à la transformation de Fourier usuelle sur $\mathbb{R}$, et nous la retrouverons au théorème 3.1 du §3.

2.2. Principe d'absorption limite. Pour tout nombre complexe ζ qui n'est pas dans le spectre $\sigma(A)$, la résolvante $R_A(\zeta) = (A-\zeta I)^{-1}$ est un opérateur borné de $L^2(\Omega)$ dans $L^2(\Omega)$. L'application $\zeta \longmapsto R_A(\zeta)$ est analytique sur $\mathbb{C}\backslash\sigma(A)$, à valeurs dans l'espace $B(L^2(\Omega), L^2(\Omega))$, mais elle n'a pas de prolongement à $\sigma(A)$: la norme de $R_A(\zeta)$ est en $|\mathrm{Im}\zeta|^{-1}$ lorsque ζ tend vers $\mu \in \sigma(A)$. En se plaçant dans un sous-espace de $L^2(\Omega)$, où la décroissance à l'infini est plus forte, on peut "absorber" la modification du comportement de la résolvante au voisinage de $\mu \in \sigma(A)$: c'est le principe d'absorption limite développé par Eidus [E69] ou Agmon [A], qui permet de donner un sens à la résolvante en des points μ du spectre. Il consiste à estimer la résolvante $R_A(\zeta)$ de manière uniforme au voisinage de $\mu \in \sigma(A)$ et à montrer l'existence d'un opérateur limite quand ζ tend vers μ.

2.2.1. Changements de variables. Pour $f, g \in L^2(\Omega)$, $\zeta \notin \sigma(A) = [c_2^2 q_1^2, +\infty)$, la formule (2.24) qui donne $(R_A(\zeta)f/g)_{L^2(\Omega)}$ est la somme des termes

$$(2.39)\quad \begin{cases} \widetilde{B}_n^i(\zeta, f, g) = \displaystyle\int_{I^i(n)} \frac{\widetilde{f}^i(\lambda, n)\overline{\widetilde{g}^i(\lambda, n)}}{\lambda - \zeta} d\lambda, \ n \in \mathbb{N}^*, \qquad i = 0, 1, \text{ou } 2, \\ I^0(n) = I_n = (c_2^2 q_n^2, c_1^2 q_n^2) \quad \text{et} \quad I^1(n) = I^2(n) = J_n = (c_1^2 q_n^2, +\infty) \end{cases}$$

où les coefficients de Fourier généralisés $(\widetilde{f}^i(\lambda, n))$ sont définis par (2.20). Les propriétés en ζ, qu'il nous faut établir pour ces termes, en vue d'obtenir le principe d'absorption limite, vont découler de théorèmes de régularité et de majorations höldériennes sur les numérateurs $\widetilde{f}^i(\lambda, n)\overline{\widetilde{g}^i(\lambda, n)}$. Nous proposons un changement de variables $\xi = \xi(\lambda)$ adapté à l'étude des termes $\widetilde{B}_n^i(\zeta, f, g)$, $i = 0, 1$, ou 2, et qui est à la base de la transformation unitaire présentée dans §2.1.3.

(i) *Changement de variables pour $i = 0$ ou 2.* Il s'écrit

$$(2.40)\quad \begin{cases} \lambda \longmapsto \xi = \xi_2(\lambda, n) = \dfrac{1}{c_2}(\lambda - c_2^2 q_n^2)^{1/2} \\ I^0(n) = (c_2^2 q_n^2, c_1^2 q_n^2) \to \mathcal{J}^0(n) = (0, a q_n) \\ I^2(n) = (c_1^2 q_n^2, +\infty) \to \mathcal{J}^2(n) = (a q_n, +\infty), \end{cases}$$

avec

$$a = \frac{1}{c_2}(c_1^2 - c_2^2)^{1/2} \left(= \frac{c_1}{c_2}(1 - \frac{c_2^2}{c_1^2})^{1/2} = \frac{c_1}{c_2} b \right).$$

L'intégrale (2.39) pour $i = 0$ ou 2 s'écrit

$$\widetilde{B}_n^i(\zeta, f, g) = \int_{\mathcal{J}^i(n)} \frac{\widetilde{f}^i(c_2^2(\xi^2 + q_n^2), n)\overline{\widetilde{g}^i(c_2^2(\xi^2 + q_n^2), n)}}{c_2^2(\xi^2 + q_n^2) - \zeta} c_2^2 2\xi d\xi.$$

Nous posons alors pour $i = 0$ et $\xi \in \mathcal{J}^0(n)$ ou pour $i = 2$ et $\xi \in \mathcal{J}^2(n)$

$$(2.41)\quad \begin{aligned} f^2(\xi, n) &= c_2\sqrt{2\xi}\ \widetilde{f}^i(c_2^2(\xi^2 + q_n^2), n) \\ &= \int_\Omega f(x, z) c_2\sqrt{2\xi}\ \varphi^i(c_2^2(\xi^2 + q_n^2), n, x) V_n(z) dx dz. \end{aligned}$$

Cette formule issue de (2.20) et (2.18) définit, pour $\xi > 0$ et $\xi \neq a q_n$, la fonction $\phi^2(\xi, n, x) = c_2\sqrt{2\xi}\ \varphi^i(c_2^2(\xi^2 + q_n^2), n, x)$. C'est une fonction propre généralisée de

l'opérateur B_n défini par (2.10), associée à la valeur propre généralisée $\lambda(\xi,n) = c_2^2(\xi^2+q_n^2)$. Nous retrouvons ainsi la fonction propre généralisée $\Psi^2(\xi,n,x,z) = \phi^2(\xi,n,x)V_n(z)$ de l'opérateur A définie par (2.27). Nous avons remarqué alors qu'elle était aussi définie pour $\xi = aq_n$ et donnée par (2.28). Compte tenu des relations

$$\begin{aligned}\xi_1' &= \xi_1'(\lambda,n) = \frac{1}{c_1}(c_1^2q_n^2-\lambda)^{1/2} = \frac{1}{c_1}[(c_1^2-c_2^2)q_n^2 - c_2^2\xi^2]^{1/2}\\ &= \frac{c_2}{c_1}(a^2q_n^2-\xi^2)^{1/2} = \frac{b}{a}(a^2q_n^2-\xi^2)^{1/2},\\ \xi_1 &= \xi_1(\lambda,n) = \frac{1}{c_1}(\lambda - c_1^2q_n^2) = \frac{c_2}{c_1}(\xi^2-a^2q_n^2)^{1/2} = \frac{b}{a}(\xi^2-a^2q_n^2)^{1/2},\end{aligned}$$

$$(c_1^2\xi_1')^2 + (c_2^2\xi_2)^2 = c_1^2c_2^2(a^2q_n^2-\xi^2) + c_2^4\xi^2,$$

$$c_1^2\xi_1 + c_2^2\xi_2 = c_1c_2(\xi^2-a^2q_n^2)^{1/2} + c_2^2\xi,$$

$$c_1^2\xi_1'(c_2^2\xi_2)^{-1} = \frac{c_1}{c_2}\frac{(a^2q_n^2-\xi^2)^{1/2}}{\xi} = \frac{a}{b}\frac{(a^2q_n^2-\xi^2)^{1/2}}{\xi},$$

nous pouvons calculer les termes (2.39) pour $i=0$ ou $i=2$.

PROPOSITION 2.3. *Soit n un entier strictement positif. Pour $\xi \in (0, aq_n]$, nous posons*

$$\phi^2(\xi,n,x) = \sqrt{\frac{2}{\pi}}\frac{c_2\xi}{[c_1^2(a^2q_n^2-\xi^2)+c_2^2\xi^2]^{1/2}}\begin{cases} exp\,[\,\frac{b}{a}(a^2q_n^2-\xi^2)^{1/2}x\,] & si\ x<0,\\ \cos(\xi x) + \frac{a}{b}\frac{(a^2q_n^2-\xi^2)^{1/2}}{\xi}\sin(\xi x) & si\ x>0.\end{cases}$$

(2.42)

Pour $\xi \in [aq_n, +\infty)$, nous posons

$$\phi^2(\xi,n,x) = \sqrt{\frac{2}{\pi}}\left[\frac{c_2^2\xi}{c_1c_2(\xi^2-a^2q_n^2)^{1/2}+c_2^2\xi}\right]^{1/2}\begin{cases}\cos[\frac{b}{a}(\xi^2-a^2q_n^2)^{1/2}x] & si\ x<0,\\ \cos(\xi x) & si\ x>0.\end{cases}$$

(2.43)

Alors la fonction $\phi^2(\xi,n,.)$ vérifie

$$(2.44)\quad B_n\phi^2(\xi,n,.) = -\frac{d}{dx}\left(c^2(x)\frac{d\phi^2}{dx}\right) + c^2(x)q_n^2\phi^2 = c_2^2(\xi^2+q_n^2)\phi^2(\xi,n,.).$$

Pour f fonction de $L^2(\Omega)$, les coefficients $f^2(\xi,n)$, déjà considérés au théorème 2.5 et définis par des intégrales convergeant au sens de $L^2((0,+\infty);d\xi)$,

$$(2.45)\quad f^2(\xi,n) = \int_\Omega f(x,z)\phi^2(\xi,n,x)V_n(z)dxdz\ \left(= \int_\Omega f(x,z)\overline{\Psi^2(\xi,n,x,z)}dxdz\right),$$

sont tels que

$$(2.46)\qquad \widetilde{B}_n^i(\zeta,f,g) = \int_{\mathcal{J}^i(n)}\frac{f^2(\xi,n)\overline{g^2(\xi,n)}}{c_2^2(\xi^2+q_n^2)-\zeta}\,d\xi, \qquad i=0 \text{ ou } 2.$$

Nous posons

$$B_n^2(\zeta,f,g)=\widetilde{B}_n^0(\zeta,f,g)+\widetilde{B}_n^2(\zeta,f,g)=\int_0^{+\infty}\frac{f^2(\xi,n)\overline{g^2(\xi,n)}}{c_2^2(\xi^2+q_n^2)-\zeta}\,d\xi. \tag{2.47}$$

(ii) *Changement de variables pour $i=1$.* Il s'écrit

$$\begin{cases}\lambda\longmapsto\xi=\xi_1(\lambda,n)=\dfrac{1}{c_1}(\lambda-c_1^2q_n^2)^{1/2}\\ I^1(n)=(c_1^2q_n^2,+\infty)\to\mathcal{J}^1(n)=(0,+\infty).\end{cases} \tag{2.48}$$

L'intégrale (2.39) pour $i=1$ s'écrit

$$\widetilde{B}_n^1(\zeta,f,g)=\int_{\mathcal{J}^1(n)}\frac{\widetilde{f}^1(c_1^2(\xi^2+q_n^2),n)\overline{\widetilde{g}^1(c_1^2(\xi^2+q_n^2),n)}}{c_1^2(\xi^2+q_n^2)-\zeta}\,c_1^2 2\xi d\xi,$$

formule qui définit pour $\xi>0$ la fonction

$$\phi^1=\phi^1(\xi,n,x)=c_1\sqrt{2\xi}\varphi^1(c_1^2(\xi^2+q_n^2),n,x).$$

Nous retrouvons ainsi la fonction propre généralisée $\Psi^1(\xi,n,x,z)=\phi^1(\xi,n,x)V_n(z)$ de l'opérateur A définie en (2.26). Compte tenu des relations

$$\begin{aligned}\xi_2&=\xi_2(\lambda,n)=\tfrac{1}{c_2}(\lambda-c_2^2q_n^2)^{1/2}=\tfrac{1}{c_2}[c_1^2\xi^2+(c_1^2-c_2^2)q_n^2]^{1/2}\\&=\frac{c_1}{c_2}(\xi^2+b^2q_n^2)^{1/2}=\frac{a}{b}(\xi^2+b^2q_n^2)^{1/2},\\ c_1^2\xi_1+c_2^2\xi_2&=c_1^2\xi_1+c_1c_2(\xi^2+b^2q_n^2)^{1/2},\\ c_1^2\xi_1(c_2^2\xi_2)^{-1}&=\frac{c_1}{c_2}\frac{\xi}{(\xi^2+b^2q_n^2)^{1/2}},\end{aligned}$$

nous pouvons calculer les termes (2.39) pour $i=1$.

PROPOSITION 2.4. *Soit n un entier strictement positif. Pour $\xi>0$, nous posons*

$$\phi^1(\xi,n,x)=\sqrt{\frac{2}{\pi}}\left[\frac{c_2(\xi^2+b^2q_n^2)^{1/2}}{c_1\xi+c_2(\xi^2+b^2q_n^2)^{1/2}}\right]^{1/2}\begin{cases}\sin(\xi x)\ \textit{si } x<0,\\ \dfrac{a}{b}\dfrac{\xi}{(\xi^2+b^2q_n^2)^{1/2}}\sin[\dfrac{a}{b}(\xi^2+b^2q_n^2)^{1/2}x]\\ \textit{si } x>0.\end{cases} \tag{2.49}$$

Alors la fonction $\phi^1(\xi,n,.)$ vérifie

$$B_n\phi^1(\xi,n,.)=-\frac{d}{dx}\left(c^2(x)\frac{d\phi^1}{dx}\right)+c^2(x)q_n^2\phi_1=c_1^2(\xi^2+q_n^2)\phi^1(\xi,n,.). \tag{2.50}$$

Pour f fonction de $L^2(\Omega)$, les coefficients $f^1(\xi,n)$, déjà considérés au théorème 2.5 et définis par des intégrales convergeant au sens de $L^2((0,+\infty);d\xi)$,

$$f^1(\xi,n)=\int_\Omega f(x,z)\phi^1(\xi,n,x)V_n(z)dxdz\ \left(=\int_\Omega f(x,z)\overline{\Psi^1(\xi,n,x,z)}dxdz\right), \tag{2.51}$$

sont tels que

$$\widetilde{B}_n^1(\zeta,f,g)=\int_{\mathcal{J}^1(n)}\frac{f^1(\xi,n)\overline{g^1(\xi,n)}}{c_1^2(\xi^2+q_n^2)-\zeta}\,d\xi\quad\textit{avec}\ \mathcal{J}^1(n)=(0,+\infty). \tag{2.52}$$

2.2.2. Définition et régularité des opérateurs de trace. Pour s réel, on note $L^2_s(\Omega)$ l'espace des fonctions mesurables à valeurs complexes définies sur Ω telles que $(1+x^2)^{s/2}f(x,z) \in L^2(\Omega)$. C'est un espace de Hilbert pour la norme

$$\|f\|_{L^2_s(\Omega)} = \|(1+x^2)^{s/2}f\|_{L^2(\Omega)},$$

qui admet $C_0^\infty(\Omega)$ comme sous-espace dense.

L'espace $B(L^2_s(\Omega),\mathbb{C})$ des formes linéaires continues sur $L^2_s(\Omega)$ est en bijection isométrique avec $L^2_{-s}(\Omega)$. Le crochet de dualité entre $L^2_s(\Omega)$ et $L^2_{-s}(\Omega)$ s'écrit pour $f \in L^2_s(\Omega)$ et $g \in L^2_{-s}(\Omega)$,

$$\langle f,g\rangle_{L^2_s(\Omega),L^2_{-s}(\Omega)} = \int_\Omega f(x,z)g(x,z)dxdz.$$

PROPOSITION 2.5. *Soit $s > 1/2$. Pour $n \geq 1, i = 1$ ou 2 et $\xi \geq 0$, on peut définir un opérateur de trace $\tau^i_n(\xi)$ continu sur $L^2_s(\Omega)$ à valeurs dans $\mathbb{C}$ tel que*

$$\forall f \in C_0^\infty(\Omega), \quad \tau^i_n(\xi)f = f^i(\xi,n) = \int_\Omega f(x,z)\phi^i(\xi,n,x)V_n(z)dxdz = \langle f, \overline{\Psi^i(\xi,n,.,.)}\rangle_{L^2_s(\Omega),L^2_{-s}(\Omega)}, \tag{2.53}$$

$$\forall f \in L^2_s(\Omega), \quad |\tau^i_n(\xi)f| \leq C|\xi|^\delta \|f\|_{L^2_s(\Omega)} \tag{2.54}$$

avec δ réel tel que $\delta \in [0,1]$ et $\delta < s - 1/2$, et avec une constante $C = C^i_n(s,\delta)$ indépendante de f et ξ. En particulier

$$\tau^1_n(0) = \tau^2_n(0) = 0. \tag{2.55}$$

La fonction $\xi \longmapsto \tau^i_n(\xi)$ définie sur $[0,+\infty]$, à valeurs dans $B(L^2_s(\Omega),\mathbb{C})$, est höldérienne en la variable ξ. Il existe $M(\xi,\xi') = M^i_n(s,\delta\,;\xi,\xi')$ fonction continue en (ξ,ξ') telle que

$$\forall f \in L^2_s(\Omega),\ |\tau^i_n(\xi)f - \tau^i_n(\xi')f| \leq M(\xi,\xi')|\xi-\xi'|^\delta \|f\|_{L^2_s(\Omega)} \tag{2.56}$$

avec δ réel tel que

(a) *$\delta \in [0,1]$ et $\delta < s - 1/2$, lorsque $i = 1$ et lorsque ξ et ξ' sont dans l'intervalle $[0,+\infty)$;*

(b) *$\delta \in [0,1]$ et $\delta < s - 1/2$, lorsque $i = 2$ et lorsque ξ et ξ' sont dans l'intervalle $[0,aq_n)$ ou ξ et ξ' dans l'intervalle $(aq_n,+\infty)$;*

(c) *$\delta \in [0,1/2]$ et $\delta < s - 1/2$, lorsque $i = 2$ et lorsque ξ et ξ' sont dans un voisinage de aq_n.*

Preuve. Il suffit de montrer les estimations (2.54) et (2.56) pour $f \in C_0^\infty(\Omega)$, sous-espace dense de $L^2_s(\Omega)$.

On établit (2.54) grâce à l'inégalité de Cauchy–Schwarz et en faisant apparaître

le poids $(1+x^2)^s$:

$$\begin{aligned}|\tau_n^i(\xi)f|^2 &= |\int_0^H \int_{-\infty}^{+\infty} f(x,z)\phi^i(\xi,n,x)V_n(z)\,dxdz|^2\\ &\le \int_0^H V_n(z)^2\,dz \int_0^H \left(\int_{-\infty}^{+\infty} |f(x,z)\phi^i(\xi,n,x)|dx\right)^2 dz\\ &\le \int_0^H \int_{-\infty}^{+\infty} |f(x,z)|^2(1+x^2)^s\,dxdz \int_{-\infty}^{+\infty} |\phi^i(\xi,n,x)|^2(1+x^2)^{-s}\,dx\\ &\le \|f\|^2_{L^2_s(\Omega)} \int_{-\infty}^{+\infty} |\phi^i(\xi,n,x)|^2(1+x^2)^{-s}\,dx.\end{aligned}$$

A partir des expressions (2.42) et (2.43) pour ϕ^2, (2.49) pour ϕ^1, on obtient (2.54). En effet, si $s > 1/2$, si $\delta \in [0,1]$ et $\delta < s - 1/2$, on a

$$\int_{-\infty}^{+\infty} \xi^2(1+x^2)^{-s}dx \le C\xi^2 \quad \text{et} \quad |\sin(\xi x)| \le \text{ Min } (|\xi x|, 1) \le |\xi x|^\delta,$$

$$\int_{-\infty}^{+\infty} |\sin(\xi x)|^2(1+x^2)^{-s}dx \le \xi^{2\delta} \int_{-\infty}^{+\infty} (1+x^2)^{\delta-s}dx \le C\xi^{2\delta}.$$

La régularité höldérienne s'obtient en estimant

$$\int_{-\infty}^{+\infty} |\phi^i(\xi,n,x) - \phi^i(\xi',n,x)|^2(1+x^2)^{-s}dx,$$

ce qui conduit à l'étude de différents termes $t(\xi,x)$ qui sont bornés sur $\overline{I\!R^+}$, analytiques sur $\overline{I\!R^+}$ si $i=1$ et sur $\overline{I\!R^+}\backslash\{aq_n\}$ si $i=2$, höldériens d'exposant $1/2$ au voisinage de $\xi = aq_n$ si $i=2$ car intervient alors la fonction racine carrée au voisinage de 0. □

Remarque 2.4. Les opérateurs $\tau_n^i(\xi)$ sont associés à la transformation unitaire du théorème 2.5. On peut aussi définir des opérateurs $\widetilde{\tau}_n^i(\lambda)$ associés à la représentation spectrale du théorème 2.4. Définis si $s > 1/2$, pour $i = 0$, 1 ou 2, $n \ge 1$ et $\lambda \in \overline{I^i(n)}$ et pour $f \in C_0^\infty(\Omega)$ par

$$\begin{aligned}(2.57)\quad \widetilde{\tau}_n^i(\lambda)f &= \widetilde{f}^i(\lambda,n)\\ &= \int_\Omega f(x,z)\overline{\psi^i(\lambda,n,x,z)}dxdz \;= \langle f, \overline{\psi^i(\lambda,n,.,.)}\rangle_{L^2_s(\Omega),L^2_{-s}(\Omega)},\end{aligned}$$

ils s'identifient par dualité à la fonction $\overline{\psi^i(\lambda,n,.,.)} = \psi^i(\lambda,n,.,.)$ qui est dans $L^2_{-s}(\Omega)$. Les changements de variables (2.40) et (2.48) donnent les relations suivantes:

$$\text{si } \lambda \in I^0(n) = (c_2^2q_n^2, c_1^2q_n^2),$$

$$(2.58)\qquad \lambda = c_2^2(\xi^2+q_n^2) \quad \text{et} \quad \tau_n^2(\xi) = c_2\sqrt{2\xi}\,\widetilde{\tau}_n^0(\lambda)\ ;$$

$$\text{si } \lambda \in I^1(n) = I^2(n) = (c_1^2q_n^2, +\infty),$$

$$(2.59)\qquad \lambda = c_1^2(\xi^2+q_n^2) \quad \text{et} \quad \tau_n^1(\xi) = c_1\sqrt{2\xi}\,\widetilde{\tau}_n^1(\lambda),$$

$$(2.60)\qquad \lambda = c_2^2(\xi^2+q_n^2) \quad \text{et} \quad \tau_n^1(\xi) = c_2\sqrt{2\xi}\,\widetilde{\tau}_n^2(\lambda);$$

$$(2.61)\qquad \widetilde{\tau}_n^0(c_2^2q_n^2) = \widetilde{\tau}_n^1(c_1^2q_n^2) = 0.$$

2.2.3. Estimations uniformes de la résolvante. Pour f et g fonctions de $L^2_s(\Omega)$, pour ζ nombre complexe n'appartenant pas au spectre $\sigma(A)$, on a

$$(R_A(\zeta)f/g)_{L^2(\Omega)} = \sum_{n\geq 1}\sum_{i=0}^{2} \widetilde{B}^i_n(\zeta, f, g).$$

Les termes

$$\widetilde{B}^i_n(\zeta, f, g) = \int_{I^i(n)} \frac{\widetilde{f^i}(\lambda, n)\overline{\widetilde{g^i}(\lambda, n)}}{\lambda - \zeta} d\lambda$$

(cf. (2.39)) ont été recalculés dans §2.2.1 à l'aide d'une intégrale en la variable $\xi \in (0, +\infty)$.

Les constantes dans les estimations qui suivent seront notées C, leur dépendance par rapport à certains paramètres n'étant spécifiée qu'en cas de besoin.

PROPOSITION 2.6. *On se donne un réel $\Lambda > 0$, $i = 1$ ou 2, et on définit les entiers*

$$N_i(\Lambda) = \text{Min}\,\{n \in \mathbb{N}^*/\Lambda \leq c_i^2 q_n^2\},\ (N_1(\Lambda) \leq N_2(\Lambda)). \tag{2.62}$$

Alors pour s réel positif, il existe une constante $C = C(\Lambda, s)$ telle que, pour toutes fonctions f et g dans $L^2_s(\Omega)$, on a

$$\sum_{n>N_1(\Lambda)}\sum_{i=1}^{2} |\widetilde{B}^i_n(\zeta, f, g)| \leq C\|f\|_{L^2_s(\Omega)}\|g\|_{L^2_s(\Omega)}, \tag{2.63a}$$

$$\sum_{n>N_2(\Lambda)}\sum_{i=0}^{2} |\widetilde{B}^i_n(\zeta, f, g)| \leq C\|f\|_{L^2_s(\Omega)}\|g\|_{L^2_s(\Omega)}, \tag{2.63b}$$

lorsque le nombre complexe ζ satisfait

$$\zeta \notin \sigma(A) \quad et \quad |\zeta| \leq \Lambda. \tag{2.64}$$

Preuve. Le dénominateur $(\lambda - \zeta)$ de la fonction à intégrer dans (2.39) est alors minoré: par exemple si $n > N_1(\Lambda)$, $|\zeta| \leq \Lambda$ et $\lambda \in I^i(n)$, $i = 1$ ou 2, on a $|\zeta - \lambda| \geq c_1^2 q_n^2 - \Lambda \geq c_1^2 q^2_{N_1(\Lambda)+1} - \Lambda = C^{-1} > 0$, d'où l'on déduit (2.63a) grâce à

$$\sum_{n\geq 1}\sum_{i=0}^{2} \int_{I^i(n)} |\widetilde{f^i}(\lambda, n)\overline{\widetilde{g^i}(\lambda, n)}| d\lambda \leq \|f\|_{L^2(\Omega)}\|g\|_{L^2(\Omega)} \leq C\|f\|_{L^2_s(\Omega)}\|g\|_{L^2_s(\Omega)}. \quad \square$$

Il reste à contrôler un nombre fini de termes dans la résolvante. Ce sont les $\widetilde{B}^i_n(\zeta, f, g)$ avec $n \leq N_1(\Lambda)$ et $i = 1$ ou 2, ou avec $n \leq N_2(\Lambda)$ et $i = 0$. Leur cardinal $N(\Lambda)$ étant fini, il suffit de faire une estimation à n et i fixés.

PROPOSITION 2.7. *On se donne un réel $\Lambda > 1$, auquel on associe les deux entiers $N_1(\Lambda)$ et $N_2(\Lambda)$ définis en (2.62). On se donne les entiers $i = 0, 1$ ou 2 et n strictement positif tels que si $i = 0$ on a $c_2^2 q_n^2 \leq \Lambda$, c'est-à-dire $n \leq N_2(\Lambda)$, ou tels que si $i = 1$ ou 2 on a $c_1^2 q_n^2 \leq \Lambda$, c'est-à-dire $n \leq N_1(\Lambda)$. Alors il existe une constante $C = C(\Lambda, s)$ telle que pour toutes fonctions f et g dans $L^2_s(\Omega)$, on a*

$$|\widetilde{B}^i_n(\zeta, f, g)| \leq C\|f\|_{L^2_s(\Omega)}\|g\|_{L^2_s(\Omega)} \tag{2.65}$$

lorsque le nombre réel s et le nombre complexe ζ satisfont

(a) $s > 1/2$, *la condition* (2.64) *et*

$$|\zeta - c_2^2 q_n^2| \geq \Lambda^{-1} \quad si\ i = 0, \qquad |\zeta - c_1^2 q_n^2| \geq \Lambda^{-1} \quad si\ i = 1; \tag{2.66}$$

(b) $s > 1$ *et la seule condition* (2.64).

Preuve. Soient i et n fixés comme indiqués. L'estimation (2.65) pour $\widetilde{B}_n^i(\zeta, f, g)$ est claire pour ζ dans un compact de $\mathbb{C}$, disjoint de $\overline{I^i(n)}$; elle doit donc être établie pour ζ variant dans un voisinage compact K de $\overline{I^i(n)}$, $\zeta \notin \overline{I^i(n)}$, et éventuellement sous la condition (2.66) qui exclut, pour $i = 0$ ou 1, un voisinage des points $c_2^2 q_n^2$ ou $c_1^2 q_n^2$.

Les changements de variables (2.40) et (2.48) de $I^i(n)$ sur $\mathcal{J}^i(n)$ peuvent être prolongés au plan complexe en posant

$$\begin{cases} Z = Z(\zeta, n) = \dfrac{1}{c_i}(\zeta - c_i^2 q_n^2)^{1/2}, & i = 0, 1 \text{ ou } 2, \quad c_0 = c_2, \\ \mathrm{Re} Z = \xi' > 0, \quad \text{ou Im } Z \geq 0 \text{ si Re } Z = 0. \end{cases} \tag{2.67}$$

Lorsque ζ décrit le voisinage compact K de $\overline{I^i(n)}$, Z décrit un voisinage compact de $\overline{\mathcal{J}^i(n)}$, et les hypothèses (2.64) et (2.66) se traduisent respectivement par

$$Z \notin \overline{I\!R^+},\ |\mathrm{Im} Z| \text{ borné et } \exists \Lambda' = \Lambda'(\Lambda) \text{ tel que } \xi' = \mathrm{Re} Z \in [0, \Lambda'], \tag{2.68}$$

et

$$\exists \Lambda'' = \Lambda''(\Lambda) > 0 \text{ tel que } \xi' = \mathrm{Re} Z > \Lambda'' \text{ si } i = 0 \text{ ou } 1. \tag{2.69}$$

Remarquons que cette condition (2.69) est aussi satisfaite si $i = 2$ et si Z est dans un voisinage assez petit de $\overline{\mathcal{J}^2(n)} = [aq_n, +\infty)$.

Lorsque $\xi \in \mathcal{J}^0(n) = (0, aq_n)$, introduisons la double notation $f^0(\xi, n) = f^2(\xi, n)$ pour le coefficient de Fourier généralisé associé à $\Psi^2(\xi, n, ., .)$(cf. la proposition 2.3). Les formules (2.46) et (2.52) s'écrivent

$$c_i^2 \widetilde{B}_n^i(\zeta, f, g) = \int_{\mathcal{J}^i(n)} \frac{f^i(\xi, n)\overline{g^i(\xi, n)}}{(\xi - Z)(\xi + Z)}\, d\xi, \qquad i = 0, 1, 2. \tag{2.70}$$

Isolons le zéro ξ' de $\mathrm{Re}(\xi - Z)$ en définissant pour $\alpha > 0$ deux intervalles complémentaires dans $\mathcal{J}^i(n)$, à savoir

$$I_\alpha^i = \{\xi \in \mathcal{J}^i(n) \,/\, |\xi - \xi'| < \alpha\} \text{ et } J_\alpha^i = \{\xi \in \mathcal{J}^i(n) \,/\, |\xi - \xi'| > \alpha\}.$$

Lorsque $\xi \in J_\alpha^i$, on a $|(\xi - Z)(\xi + Z)| \geq \alpha^2$ et

$$\left| \int_{J_\alpha^i} \frac{f^i(\xi, n)\overline{g^i(\xi, n)}}{(\xi - Z)(\xi + Z)}\, d\xi \right| \leq C(\alpha, s) \|f\|_{L_s^2(\Omega)}\, \|g\|_{L_s^2(\Omega)} \quad \text{pour } s \geq 0.$$

La proposition sera donc établie si nous montrons une estimation du type (2.65) pour

$$c_\alpha^i(\zeta, f, g) = \int_{I_\alpha^i} \frac{f^i(\xi, n)\overline{g^i(\xi, n)}}{(\xi - Z)(\xi + Z)}\, d\xi.$$

Ce sont les propriétés des traces $\tau_n^i(\xi)f = f^i(\xi,n)$ établies à la proposition 2.5 qui vont la donner. Pour cela, nous décomposons le numérateur en posant

$$h^i(\xi,f,g) = \begin{cases} 0 \text{ si } \xi \notin \overline{\mathcal{J}^i(n)}, \\ f^i(\xi,n)\overline{g^i(\xi,n)} \text{ si } \xi \in \mathcal{J}^i(n) \end{cases} \tag{2.71}$$

$$= h^i(\xi',f,g) + (f^i(\xi,n) - f^i(\xi',n))\overline{g^i(\xi,n)} + f^i(\xi',n)(\overline{g^i(\xi,n)} - \overline{g^i(\xi',n)}).$$

Nous estimons d'abord

$$d_\alpha^i(\xi') = \int_{I_\alpha^i} \frac{h^i(\xi',f,g)}{(\xi - Z)(\xi + Z)} d\xi.$$

Pour $\xi' \notin \mathcal{J}^i(n)$, on a $d_\alpha^i(\xi') = 0$. Pour $\xi' \in \mathcal{J}^i(n)(\subset (0,+\infty))$, on a

$$d_\alpha^i(\xi') = h^i(\xi',f,g) \int_{I_\alpha^i} \frac{d\xi}{(\xi - Z)(\xi + Z)}$$

avec $|h^i(\xi',f,g) \leq C\xi'^{2\delta}\|f\|_{L^2_s(\Omega)}\ \|g\|_{L^2_s(\Omega)}$ si $\delta \in [0,1]$, $\delta < s - 1/2$, grâce à (2.54). L'intégrale de

$$\frac{1}{(\xi - Z)(\xi + Z)} = \frac{1}{2Z}\left(\frac{1}{\xi - Z} - \frac{1}{\xi + Z}\right)$$

doit être évaluée avec précision. Pour $\xi' \in \mathcal{J}^i(n)$ et $Z = \xi' + i\ \mathrm{Im}\, Z$ vérifiant (2.68), on a

$$\left|\frac{1}{2Z}\int_{I_\alpha^i} \frac{d\xi}{\xi \pm Z}\right| = \left|\frac{1}{2Z}\left[\frac{1}{2}\log\{(\xi \pm \xi')^2 + \mathrm{Im}\, Z^2\} \mp i \arctan\left(\frac{\xi \pm \xi'}{\mathrm{Im} Z}\right)\right]_{\mathrm{Min} I_\alpha^i}^{\mathrm{Max} I_\alpha^i}\right|$$

$\leq C(\xi')^{-1-\gamma}$ avec γ réel strictement positif arbitraire. Introduisant les conditions (2.68) et (2.69), on obtient:

(i) $|d_\alpha^i(\xi')| \leq C\|f\|_{L^2_s(\Omega)}\|g\|_{L^2_s(\Omega)}$ pour $s > 1/2$, si $\xi' \in [\Lambda'', \Lambda']$, soit encore, lorsque ζ satisfait (2.64) si $i = 2$, et lorsque ζ satisfait (2.64) et (2.66) si $i = 0$ ou 1,

(ii) $|d_\alpha^i(\xi')| \leq C\|f\|_{L^2_s(\Omega)}\|g\|_{L^2_s(\Omega)}$ pour $s > 1$ si $\xi' \in [0, \Lambda']$, soit encore lorsque ζ satisfait simplement (2.64) si $i = 0$ ou 1, puisqu'il faut alors choisir δ et γ tels que $2\delta - 1 - \gamma \geq 0$ (on utilise (2.54) avec $\delta > 1/2$).

Estimons enfin

$$e_\alpha^i(\xi') = c_\alpha^i(\zeta,f,g) - d_\alpha^i(\xi') = \int_{I_\alpha^i} \frac{h^i(\xi,f,g) - h^i(\xi',f,g)}{(\xi - Z)(\xi + Z)} d\xi.$$

La décomposition (2.71) de $(h^i(\xi,f,g) - h^i(\xi',f,g))$ et les estimations (2.54) et (2.56) des opérateurs de trace donnent

$$|h^i(\xi,f,g) - h^i(\xi',f,g)| \leq M|\xi - \xi'|^{\delta+\varepsilon}(\xi^\delta + \xi'^\delta)\|f\|_{L^2_s(\Omega)}\ \|g\|_{L^2_s(\Omega)},$$

avec $\delta \geq 0$ et $\varepsilon > 0$ fixés tels que

$$\begin{cases} \delta + \varepsilon < \mathrm{Min}(1, s - 1/2) \text{ si } (i = 1) \text{ ou bien si } (\,(i = 0 \text{ ou } 2) \text{ et } |\xi' - aq_n| > \Lambda''), \\ \delta + \varepsilon < \mathrm{Min}(1/2, s - 1/2) \text{ si } i = 0 \text{ ou } 2, \end{cases}$$

avec la constante $M = M(s, \delta, \varepsilon, \Lambda')$ ou $M(s, \delta, \varepsilon, \Lambda', \Lambda'')$ qui est indépendante d'une part de ξ et ξ' dans $[0, \Lambda']$ ou $[\Lambda'', \Lambda']$, et d'autre part de f et g dans $L^2_s(\Omega)$.

Les minorations $|\xi - Z| \geq |\mathrm{Re}(\xi - Z)| = |\xi - \xi'|$ et $|\xi + Z| \geq \xi + \xi'$ entraînent alors

$$|e^i_\alpha(\xi')| \leq M \|f\|_{L^2_s(\Omega)} \, \|g\|_{L^2_s(\Omega)} \int_{I^i_\alpha} \frac{c(\xi, \xi')}{|\xi - \xi'|^{1-\varepsilon}} d\xi \ .$$

Des majorations uniformes, quand $\xi \in I^i_\alpha$, $\xi' \in [\Lambda'', \Lambda']$ ou $\xi' \in [0, \Lambda']$, de la fonction à intégrer

$$c(\xi, \xi') = \frac{|\xi - \xi'|^\delta (\xi^\delta + \xi'^\delta)}{\xi + \xi'}$$

terminent la preuve de la proposition 2.7. □

Ces propositions 2.6 et 2.7 permettent de majorer uniformément la norme de la résolvante dans l'espace $B(L^2_s(\Omega), L^2_{-s}(\Omega))$.

THÉORÈME 2.6. *On se donne un réel $\Lambda > 1$.*

(a) *Soit $s > 1$. Alors il existe une constante $C = C(\Lambda, s)$ telle que pour toute fonction f dans $L^2_s(\Omega)$ et pour tout nombre complexe ζ satisfaisant*

$$\zeta \notin \sigma(A) \quad et \quad |\zeta| \leq \Lambda, \tag{2.72}$$

on a $\|R_A(\zeta) f\|_{L^2_{-s}(\Omega)} \leq C \|f\|_{L^2_s(\Omega)}$.

(b) *Soit $s > 1/2$. Alors il existe une constante $C = C(\Lambda, s)$ telle que pour toute fonction f dans $L^2_s(\Omega)$ et pour tout nombre complexe ζ satisfaisant (2.72) et*

$$|\zeta - c_1^2 q_n^2| \geq \Lambda^{-1} \quad et \quad |\zeta - c_2^2 q_n^2| \geq \Lambda^{-1} \quad pour\ n \in \mathbb{N}^*, \tag{2.73}$$

on a $\|R_A(\zeta) f\|_{L^2_{-s}(\Omega)} \leq C \|f\|_{L^2_s(\Omega)}$.

Remarque 2.5. Dans le cas $c_1 = c_2$, la preuve du théorème 2.6 est inopérante. En effet, pour $\xi = 0$, les fonctions (2.38) ne sont pas identiquement nulles, égales à

$$\Psi^2(0, n, x, z) = \phi^2(0, n, x) V_n(z) = \frac{1}{\sqrt{\pi}} V_n(z). \tag{2.74}$$

Il s'ensuit que le noyau de l'opérateur $\tau^2_n(0)$ (cf. la proposition 2.5) est un hyperplan fermé de $L^2_s(\Omega)$ défini par

$$NL^2_s(n) = \left\{ f \in L^2_s(\Omega) \ / \ \int_\Omega f(x, z) \, V_n(z) \, dxdz = 0 \right\}. \tag{2.75}$$

Nous montrerons au théorème 3.3, que l'estimation fondamentale (2.65) sous les conditions (2.64), n'a pas lieu. Précisément, nous construirons f_n, fonction dans $\bigcap_{s \geq 0} L^2_s(\Omega)$, telle que

$$\lim_{\zeta \to \lambda(0,n) \, , \, \zeta \notin \sigma(A)} |\widetilde{B}^2_n(\zeta, f_n, f_n)| \ = \ +\infty. \tag{2.76}$$

COROLLAIRE 2.4. *Sous les conditions du théorème 2.6, on a*

$$C_1 \|\nabla R_A(\zeta) f\|_{(L^2_{-s}(\Omega))^2} \leq \|R_A(\zeta) f\|_{L^2_{-s}(\Omega)} + \|A R_A(\zeta) f\|_{L^2_{-s}(\Omega)} \leq C_2 \|f\|_{L^2_s(\Omega)}. \tag{2.77}$$

Remarque 2.6. Le domaine $D(A)$ de l'opérateur $A = -\partial_x(c^2(x)\partial_x) - c^2(x)\partial_z^2$, défini par (2.1) et (2.2), n'est pas inclus dans $H^2(\Omega)$. La norme du graphe sur $D(A)$ ne peut donc être comparée à la norme naturelle de $H^2(\Omega)$, comme cela est le cas pour l'opérateur $-\Delta_c = -c^2(x)(\partial_x^2 + \partial_z^2)$, de domaine $\{u \in H^2(\Omega) \ / \ \partial_z u|_{z=0} = u|_{z=H} = 0\}$, qui est auto-adjoint positif dans $L^2(\Omega; c^{-2}(x)dxdz)$.

2.2.4. Résultats de convergence. A partir des propositions 2.5, 2.6, et 2.7, on établit les convergences et estimations suivantes au voisinage de $\mu \in \sigma(A)$.

PROPOSITION 2.8. *Soient $s \geq 0$, f et $g \in L^2_s(\Omega)$. Soient $\mu \in \sigma(A)$ et $I\!N(\mu) = \{(i,n) \ / \ i = 0,1 \text{ ou } 2\ , n \in I\!N^*,\ \mu \notin \overline{I^i(n)}\}$. Alors la limite quand ζ tend vers μ de*

$$S_{I\!N(\mu)}(\zeta, f, g) = \sum_{(i,n)\in I\!N(\mu)} \widetilde{B}^i_n(\zeta, f, g), \tag{2.78}$$

existe et est égale à

$$S_{I\!N(\mu)}(\mu, f, g) = \sum_{(i,n)\in I\!N(\mu)} \widetilde{B}^i_n(\mu, f, g), \tag{2.79}$$

avec

$$\widetilde{B}^i_n(\mu, f, g) = \int_{I^i(n)} \frac{\widetilde{f}^i(\lambda, n)\overline{\widetilde{g}^i(\lambda, n)}}{\lambda - \mu} d\lambda,\ (i,n) \in I\!N(\mu). \tag{2.80}$$

De plus, pour tout compact $K \subset \mathbb{C}$, il existe une constante $C = C(s,K)$ telle que si

$$I\!N(K) = \{(i,n) \ / \ K \cap \overline{I^i(n)} = \emptyset\} \quad et \quad S_{I\!N(K)}(\zeta, f, g) = \sum_{(i,n)\in I\!N(K)} \widetilde{B}^i_n(\zeta, f, g),$$

alors

$$\forall f, g \in L^2_s(\Omega),\ \forall \zeta \in K,\ |\ S_{I\!N(K)}(\zeta, f, g)| \leq C \|f\|_{L^2_s(\Omega)} \|g\|_{L^2_s(\Omega)}. \tag{2.81}$$

Preuve. On procède comme dans la preuve de la proposition 2.6 et on applique le théorème de la convergence dominée. □

PROPOSITION 2.9. *Soient $s > 1/2$, f et $g \in L^2_s(\Omega)$. Soient $i = 0,1$ ou 2 et n entier strictement positif fixés. Alors les limites suivantes existent et vérifient*

(a) *Si $\mu \in I^i(n)$,*

$$\begin{cases} \widetilde{B}^{i,\pm}_n(\mu, f, g) = \lim\limits_{\zeta \to \mu, \pm \mathrm{Im}\zeta > 0} \widetilde{B}^i_n(\zeta, f, g) \\ = v.p. \left(\displaystyle\int_{I^i(n)} \frac{\widetilde{f}^i(\lambda, n)\overline{\widetilde{g}^i(\lambda, n)}}{\lambda - \mu} d\lambda \right) \pm i\pi \widetilde{f}^i(\mu, n)\overline{\widetilde{g}^i(\mu, n)}. \end{cases} \tag{2.82}$$

De plus, pour tout compact K de $\mathbb{C}^{\pm} = \{\zeta \in \mathbb{C} / \pm \mathrm{Im}\zeta \geq 0\}$ qui ne contient pas les extrémités de $I^i(n)$, on a l'existence d'une constante $C = C(s,K)$ telle que

$$\forall f, g \in L^2_s(\Omega), \qquad \forall \zeta \in K, \quad |\widetilde{B}^{i,\pm}_n(\zeta, f, g)| \leq C \|f\|_{L^2_s(\Omega)} \|g\|_{L^2_s(\Omega)} \tag{2.83}$$

où on a posé $\widetilde{B}^{i,\pm}_n(\zeta, f, g) = \widetilde{B}^i_n(\zeta, f, g)$ si $\zeta \notin \overline{I^i(n)}$.

(b) *Si $\mu = c_1^2 q_n^2$,*

$$\begin{cases} B^{2,\pm}_n(c_1^2 q_n^2, f, g) = \lim\limits_{\zeta \to c_1^2 q_n^2, \pm \mathrm{Im}\zeta > 0} B^2_n(\zeta, f, g) \\ = v.p. \left(\displaystyle\int_0^{+\infty} \frac{f^2(\xi, n)\overline{g^2(\xi, n)}}{c_2^2(\xi^2 + q_n^2) - c_1^2 q_n^2} d\xi \right) \pm i\pi \widetilde{f}^0(c_1^2 q_n^2, n)\overline{\widetilde{g}^0(c_1^2 q_n^2, n)}. \end{cases} \tag{2.84}$$

On rappelle que $B_n^2(\zeta, f, g)$ *a été défini en* (2.47). *De plus, pour tout compact* K *de* $\mathbb{C}^{\pm}$ *qui ne contient pas l'extrémité* $c_2^2 q_n^2$ *de* $I^0(n)$, *on a l'existence d'une constante* $C = C(s, K)$ *telle que*

$$\forall f, g \in L_s^2(\Omega),\ \forall \zeta \in K,\ |B_n^{2,\pm}(\zeta, f, g)| \leq C \|f\|_{L_s^2(\Omega)} \|g\|_{L_s^2(\Omega)} \tag{2.85}$$

où on a posé $B_n^{2,\pm}(\zeta, f, g) = B_n^2(\zeta, f, g)$ *si* $\zeta \notin [c_2^2 q_n^2, +\infty)$.

Preuve. On utilise les propriétés höldériennes des opérateurs de trace énoncées à la proposition 2.5, l'égalité

$$(\lambda - \mu \pm i0)^{-1} = v.p.(\frac{1}{\lambda - \mu}) \mp i\pi\delta_\mu$$

au sens des distributions, la formule (2.41) exprimant f^2 en ξ en fonction de $\widetilde{f}^0$ ou $\widetilde{f}^2$ en $\lambda = c_2^2(\xi^2 + q_n^2)$ ou encore les formules (2.58) et (2.59) reliant les opérateurs de trace correspondants, et enfin l'estimation (2.65) de la proposition 2.7. □

PROPOSITION 2.10. *Soient* $s > 1$, $f, g \in L_s^2(\Omega)$, *et* n *entier strictement positif. Alors les limites suivantes existent et vérifient*

$$\begin{cases} \widetilde{B}_n^0(c_2^2 q_n^2, f, g) = \lim\limits_{\zeta \to c_2^2 q_n^2, \zeta \notin \sigma(A)} \widetilde{B}_n^0(\zeta, f, g) \\ = \displaystyle\int_{I^0(n)} \frac{\widetilde{f}^0(\lambda, n)\overline{\widetilde{g}^0(\lambda, n)}}{\lambda - c_2^2 q_n^2} d\lambda = \int_0^{aq_n} \frac{f^2(\xi, n)\overline{g^2(\xi, n)}}{c_2^2 \xi^2} d\xi, \end{cases} \tag{2.86}$$

$$\begin{cases} \widetilde{B}_n^1(c_1^2 q_n^2, f, g) = \lim\limits_{\zeta \to c_1^2 q_n^2, \zeta \notin \sigma(A)} \widetilde{B}_n^1(\zeta, f, g) \\ = \displaystyle\int_{I^1(n)} \frac{\widetilde{f}^1(\lambda, n)\overline{\widetilde{g}^1(\lambda, n)}}{\lambda - c_1^2 q_n^2} d\lambda = \int_0^{+\infty} \frac{f^1(\xi, n)\overline{g^1(\xi, n)}}{c_1^2 \xi^2} d\xi. \end{cases} \tag{2.87}$$

De plus, pour tout compact $K \subset \mathbb{C}^{\pm} = \{\zeta \in \mathbb{C} / \pm \mathrm{Im}\zeta \geq 0\}$, *on a l'existence d' une constante* $C = C(s, K)$ *telle que, pour* $i = 0$ *ou* 1,

$$\forall f, g \in L_s^2(\Omega),\ \forall \zeta \in K,\ |\widetilde{B}_n^{i,\pm}(\zeta, f, g)| \leq C \|f\|_{L_s^2(\Omega)} \|g\|_{L_s^2(\Omega)} \tag{2.88}$$

où on a posé $\widetilde{B}_n^{i,\pm}(\zeta, f, g) = \widetilde{B}_n^i(\zeta, f, g)$ *si* $\zeta \notin \overline{I^i(n)}$, *ou bien si* $i = 0$ *et* $\zeta = c_2^2 q_n^2$, *ou bien si* $i = 1$ *et* $\zeta = c_1^2 q_n^2$.

Preuve. On utilise les propositions 2.5 et 2.7, la formule (2.46) si $i = 0$, la formule (2.52) si $i = 1$. □

On peut maintenant énoncer le principe d'absorption limite pour A. Pour le détail de la preuve, on prend la même démarche que dans [A] et [DG86].

THÉORÈME 2.7 (Principe d'absorption limite faible). *On considère la résolvante* $\zeta \mapsto R_A(\zeta)$ *comme une fonction définie sur* $\mathbb{C} \backslash \sigma(A)$ *à valeurs dans* $B(L_s^2(\Omega), L_{-s}^2(\Omega))$ *avec* s *réel positif. Soient* f, g *fonctions dans* $L_s^2(\Omega)$, *et* μ, s *réels tels que*

$$\begin{cases} \mu \in \sigma(A) \text{ et } s > 1, \\ \text{ou} \\ \mu \in \sigma(A) \setminus \{c_2^2 q_n^2, c_1^2 q_n^2 \ / \ n \in \mathbb{N}^*\} \text{ et } s > 1/2. \end{cases} \tag{2.89}$$

(i) *Les deux limites suivantes existent pour la topologie de la convergence faible dans l'espace* $L_{-s}^2(\Omega)$:

$$R_A^{\pm}(\mu) f = \lim_{\zeta \to \mu,\ \pm \mathrm{Im}\zeta > 0} R_A(\zeta) f. \tag{2.90}$$

(ii) *Les fonctions* $u^{\pm} = R_A^{\pm}(\mu)f$ *vérifient au sens des distributions*

$$(A - \mu I)u^{\pm} = f.$$

(iii) *Avec les notations des propositions* 2.9 *et* 2.10, *on a*
(iiia) *Si* $\mu \notin \{c_1^2 q_n^2 / n \in I\!N^*\}$,

$$\langle R_A^{\pm}(\mu)f, g\rangle_{L^2_{-s}(\Omega)\,,\,L^2_s(\Omega)} = \sum_{n\geq 1}\sum_{i=0,1,2} \widetilde{B}_n^{i,\pm}(\mu, f, \overline{g}). \tag{2.91}$$

(iiib) *Si* $\mu = c_1^2 q_m^2$ *et donc* $s > 1$,

$$\langle R_A^{\pm}(c_1 q_m^2)f, g\rangle_{L^2_{-s}(\Omega)\,,\,L^2_s(\Omega)} \tag{2.92}$$

$$= \sum_{n\geq 1, n\neq m}\sum_{i=0,1,2} \widetilde{B}_n^{i,\pm}(c_1^2 q_m^2, f, \overline{g}) + B_m^{2,\pm}(c_1^2 q_m^2, f, \overline{g}) + \widetilde{B}_m^1(c_1^2 q_m^2, f, \overline{g}).$$

Remarque 2.7. On a

$$\widetilde{B}_n^{i,+}(\mu, f, g) = \widetilde{B}_n^{i,-}(\mu, f, g) \text{ si } \mu \notin \overline{I^i(n)} \text{ ou bien si } \widetilde{f}^i(\mu, n)\widetilde{g}^i(\mu, n) = 0,$$

$$B_n^{2,+}(c_1^2 q_n^2, f, g) = B_n^{2,-}(c_1^2 q_n^2, f, g) \text{ si } \widetilde{f}^0(c_1^2 q_n^2, n)\widetilde{g}^0(c_1^2 q_n^2, n) = 0.$$

THÉORÈME 2.8 (Principe d'absorption limite). *On considère la résolvante* $\zeta \mapsto R_A(\zeta)$ *comme une fonction définie sur* $\mathbb{C}\backslash\sigma(A)$ *à valeurs dans* $B(L^2_s(\Omega), L^2_{-s}(\Omega))$. *Alors, sous la condition* (2.89), *les deux limites suivantes existent pour la topologie de la norme dans* $B(L^2_s(\Omega), L^2_{-s}(\Omega))$:

$$R_A^{\pm}(\mu) = \lim_{\zeta\to\mu,\ \pm \mathrm{Im}\zeta > 0} R_A(\zeta). \tag{2.93}$$

Remarque 2.8. On peut vérifier que la fonction $\zeta \mapsto R_A^{\pm}(\zeta)$ ainsi prolongée est localement höldérienne sur $\mathbb{C}^{\pm} = \{\zeta \in \mathbb{C} \ / \ \pm \mathrm{Im}\zeta \geq 0\}$. En particulier sa norme est bornée lorsque le module $|\ \zeta\ |$ reste borné et $s > 1$. On retrouve ainsi un résultat connu pour le Laplacien dans tout l'espace (cf. [E69] et [W86] pour des résultats voisins), ainsi que pour certains guides d'ondes (cf. [MW88]).

Remarque 2.9. Dans le cas de la bande homogène $c_1 = c_2 = c$, et pour des conditions limites de Dirichlet ou de Neumann, la violation du principe d'absorption limite au voisinage des seuils a été établie par Werner dans [W87]. Nous considérons ce cas dans la section 3 qui suit. Nous verrons que les théorèmes 2.7 et 2.8 sont alors non valides au voisinage des seuils, et qu'il est nécessaire de se placer sur un hyperplan fermé de $L^2_s(\Omega)$ pour obtenir un principe d'absorption limite.

3. Profil de vitesse c(x,z) = c(z) minoré par $c_m > 0$. Nous travaillons, dans cette section, avec un profil de vitesse indépendant de la variable x et satisfaisant l'hypothèse

(H) $\qquad c \in L^\infty((0, H)) \quad \text{et} \quad \mathrm{Min}\, c(z) \geq c_m > 0.$

Le profil (1.6) en est un cas particulier.

Si l'étude menée est similaire à celle du §2, les résultats sont différents et nous développons particulièrement les points spécifiques à ce profil $c(x, z) = c(z)$: il s'agit

de l'étude des opérateurs de trace et des estimations de la résolvante au voisinage des seuils $\lambda_n, n \geq 1$, qui sont ici les valeurs propres d'un opérateur réduit B_0, associé à l'opérateur A par transformation de Fourier en la variable x.

On montre au théorème 3.3 que la résolvante "explose" au voisinage de ces seuils au sens où, pour tout $s \geq 0$, il existe f_n dans $L^2_s(\Omega)$ telle que:

$$\lim_{\zeta \to \lambda_n} \|R_A(\zeta) f_n\|_{L^2_{-s}(\Omega)} = +\infty,$$

si bien que le principe d'absorption limite ne peut être établi en λ_n que sur un sous-espace strict $NL^2_s(n)$ de $L^2_s(\Omega)$.

3.1. Théorie spectrale de l'opérateur $A = -\nabla \cdot (c^2(z)\nabla)$. On trouve certains des résultats qui suivent dans le chapitre II §10 de [Wi2].

L'opérateur auto-adjoint positif dans $L^2(\Omega)$ est maintenant

$$A = -c^2(z)\partial_x^2 - \partial_z(c^2(z)\partial_z), \tag{3.1}$$

de domaine

$$D(A) = \{u \in H^1(\Omega) \ / \ Au \in L^2(\Omega) \ \text{ et } \ (c^2\partial_z u)|_{z=0} = u|_{z=H} = 0\}. \tag{3.2}$$

Comme il a été dit dans l'introduction, ces conditions limites sont celles du problème physique posé, et nous pourrions choisir d'autres conditions (CL), Dirichlet ou Neumann en $z = 0$ ou $z = H$ (cf. aussi la remarque 4.2).

La séparation des variables x et z dans l'espace $L^2(\Omega; dxdz) = L^2(\mathbb{R}; dx) \otimes L^2((0,H); dz)$ permet de ramener l'analyse spectrale de A, opérateur à coefficients indépendants de x, à celle d'une famille d'opérateurs B_ξ, ξ réel, auto-adjoints à inverse compact sur $L^2((0,H); dz)$.

Plus précisément, nous utilisons l'isomorphisme de Fourier défini sur $L^2(\mathbb{R}; dx)$ par

$$v \in L^2(\mathbb{R}; dx) \ , \ \hat{v}(\xi) = (2\pi)^{-1/2} \int_{\mathbb{R}} v(x) e^{-i\xi x} dx. \tag{3.3}$$

Remarquons que cela revient à choisir les fonctions

$$(\ \varphi(\xi, .) : x \longmapsto (2\pi)^{-1/2} e^{i\xi x}\ ), \ \xi \in \mathbb{R}, \tag{3.4}$$

comme famille complète de fonctions propres généralisées pour l'opérateur $A_1 = -\dfrac{d^2}{dx^2}$, de domaine $D(A_1) = \{v \in H^1(\mathbb{R}) \ / \ A_1 v \in L^2(\mathbb{R})\} = H^2(\mathbb{R})$, auto-adjoint dans $L^2(\mathbb{R})$.

PROPOSITION 3.1. *L'espace de Hilbert $L^2(\Omega; dxdz)$ et l'opérateur A sont respectivement unitairement équivalents à*

$$\begin{cases} \widetilde{\mathcal{H}} = L^2(\Omega; d\xi dz) = \displaystyle\int_{\mathbb{R}}^{\oplus} H_\xi d\xi, \ \textit{intégrale directe des espaces} \\ H_\xi = L^2((0,H); dz), \ \xi \in \mathbb{R}, \end{cases} \tag{3.5}$$

et

$$\begin{cases} \widetilde{A} = \displaystyle\int_{\mathbb{R}}^{\oplus} B_\xi d\xi, \ \textit{intégrale directe des opérateurs non bornés} \\ B_\xi = -\dfrac{d}{dz}\left(c^2(z)\dfrac{d}{dz}\right) + c^2(z)\xi^2, \ \xi \in \mathbb{R}, \end{cases} \tag{3.6}$$

de domaines $D(B_\xi)$, *indépendants du réel* ξ, *donnés par*

$$D(B) = \{u \in H^1((0,H)) \,/\, c^2u' \in H^1((0,H)) \;\; et \;\; (c^2u')(0) = u(H) = 0\}. \tag{3.7}$$

Avec la définition (3.4) *de* $\varphi(\xi,.)$ *on a, pour toute fonction* u *dans* $D(A)$,

$$u(x,z) = \int_{\mathbb{R}} \hat{u}(\xi,z)\varphi(\xi,x)d\xi \;\; et \;\; Au(x,z) = \int_{\mathbb{R}} B_\xi \hat{u}(\xi,z)\varphi(\xi,x)d\xi, \tag{3.8}$$

la convergence de ces intégrales ayant lieu dans $L^2(\Omega; dxdz)$.

Pour ξ réel fixé, B_ξ est un opérateur auto-adjoint dans $L^2((0,H))$, positif à inverse compact. Son spectre $\sigma(B_\xi)$ est minoré par $\xi^2 \operatorname{Min} c^2(z)$, purement ponctuel et dénombrable, soit

$$\sigma(B_\xi) = \{\lambda(\xi,n) \,/\, n \in \mathbb{N}^*\}, \tag{3.9}$$

Il est montré au théorème 4.5 de l'annexe que les valeurs propres $\lambda(\xi,n)$ sont simples, et on convient que la suite $(\lambda(\xi,n))_{n\geq 1}$ est (strictement) croissante.

L'espace $L^2((0,H))$ admet donc une base orthonormée $(V(\xi,n,.))_{n\geq 1}$ de fonctions propres de B_ξ. Ces fonctions feuvent être choisies rielles et satisfont

$$V(\xi,n,.) \in D(B) \;\; \text{et} \;\; B_\xi(V(\xi,n,.)) = \lambda(\xi,n)V(\xi,n,.), \tag{3.10a}$$

$$\|V(\xi,n,.)\|_{L^2((0,H))} = \left(\int_0^H |V(\xi,n,z)|^2 dz\right)^{1/2} = 1. \tag{3.10b}$$

Remarquons que lorsque $c_1 = c_2 = c$, on a $B_\xi = c^2A_2 + c^2\xi^2$, où A_2 est l'opérateur (2.8). Selon les formules (2.9) ou (4.36), on a alors $\lambda(\xi,n) = c^2q_n^2 + c^2\xi^2$ et on peut choisir $V(\xi,n,z) = V_n(z)$.

La transformation unitaire

$$\widetilde{\mathcal{F}}_\xi \begin{cases} L^2((0,H)) \longrightarrow \widetilde{\mathcal{H}}_\xi = \bigoplus_{n\geq 1} \mathbb{C}V(\xi,n,.) \\ v \longmapsto \sum_{n\geq 1} (v/V(\xi,n,.))_{L^2((0,H))} V(\xi,n,.) \end{cases} \tag{3.11}$$

définit une représentation spectrale de B_ξ.

THÉORÈME 3.1 (Famille complète de fonctions propres généralisés pour A et transformation unitaire associée). *Les fonctions* $\Psi = \Psi(\xi,n,x,z)$, $\xi \in \mathbb{R}$, $n \in \mathbb{N}^*$, *définies sur* Ω *à partir de* (3.4) *et* (3.10) *par*

$$\Psi(\xi,n,x,z) = \varphi(\xi,x)V(\xi,n,z) = (2\pi)^{-1/2} e^{i\xi x} V(\xi,n,z), \tag{3.12}$$

sont des fonctions propres généralisées de A. *Elles satisfont l'équation différentielle* $A\Psi(\xi,n,.,.) = \lambda(\xi,n)\Psi(\xi,n,.,.)$. *Elles forment une famille complète au sens où elles permettent de définir la transformation unitaire*

$$\widetilde{\mathcal{F}} \begin{cases} L^2(\Omega) \longrightarrow \widetilde{\mathcal{H}} = \int_{\mathbb{R}}^{\oplus} \widetilde{\mathcal{H}}_\xi d\xi \\ f \longmapsto \widetilde{\mathcal{F}}f \;\; avec \;\; \widetilde{\mathcal{F}}f(\xi) = \widetilde{\mathcal{F}}_\xi(\hat{f}(\xi,.)), \; \xi \in \mathbb{R}. \end{cases} \tag{3.13}$$

Les coefficients associés à f fonction de $L^2(\Omega)$, pour presque tout ξ réel et pour n entier strictement positif, s'écrivent

$$\tilde{f}(\xi,n) = (\tilde{\mathcal{F}}_\xi \hat{f}(\xi,.)/V(\xi,n,.))_{L^2((0,H))} = \int_0^H \hat{f}(\xi,z)V(\xi,n,z)dz \tag{3.14}$$

$$= L^2(\mathbb{R};d\xi) - \lim_{X\to+\infty} \int_{\{(x,z)\in\Omega/|x|<X\}} f(x,z)\overline{\Psi(\xi,n,x,z)}dxdz.$$

Si f appartient au domaine $D(A)$, on a la formule de diagonalisation

$$\widetilde{Af}(\xi,n) = \lambda(\xi,n)\tilde{f}(\xi,n). \tag{3.15}$$

L'égalité de Bessel–Parseval pour toutes fonctions f et g de $L^2(\Omega)$ s'écrit

$$(f \mid g)_{L^2(\Omega)} = \sum_{n\geq 1} \int_{\mathbb{R}} \tilde{f}(\xi,n)\overline{\tilde{g}(\xi,n)}\, d\xi. \tag{3.16}$$

La reconstruction d'une fonction f de $L^2(\Omega)$ à partir de ses coefficients de Fourier généralisés s'écrit

$$f(x,z) = \sum_{n\geq 1} \int_{\mathbb{R}} \tilde{f}(\xi,n)\Psi(\xi,n,x,z)\, d\xi, \tag{3.17}$$

les convergences étant toujours à entendre au sens L^2.

COROLLAIRE 3.1. *Le spectre $\sigma(A)$ et la résolvante $R_A(\zeta) = (A-\zeta I)^{-1}$ vérifient*

$$\sigma(A) = \overline{\bigcup_{\xi\in\mathbb{R}} \sigma(B_\xi)} = \overline{\{\lambda(\xi,n) \,/\, \xi\in\mathbb{R} \text{ et } n\in\mathbb{N}^*\}} = [\lambda(0,1),+\infty), \tag{3.18}$$

$$\forall\ \zeta\notin\sigma(A),\ \forall f,g\in L^2(\Omega), (R_A(\zeta)f/g)_{L^2(\Omega)} = \sum_{n\geq 1}\int_{\mathbb{R}} \frac{\tilde{f}(\xi,n)\overline{\tilde{g}(\xi,n)}}{\lambda(\xi,n)-\zeta}\, d\xi. \tag{3.19}$$

Dans le théorème 3.1, les fonctions et valeurs propres généralisées de A sont paramétrées par $n\in\mathbb{N}^*$ et $\xi\in\mathbb{R}$. On montre, au théorème 4.5 de l'annexe, que les fonctions $\xi\mapsto\lambda(\xi,n)$ sont paires, strictement croissantes sur $\mathbb{R}^+$ de $\lambda(0,n)$ à $+\infty$, et analytiques. On peut revenir à la variable spectrale λ de A par le changement de variables

$$(\lambda = \lambda(.,n) \ :\ \xi \longmapsto \lambda(\xi,n)\in I_n = (\lambda(0,n),+\infty)) \tag{3.20}$$

qui est un difféomorphisme analytique de $(0,+\infty)$ ou $(-\infty,0)$ sur I_n, d'inverses ($\xi^+ = \xi^+(.,n)$: $\lambda\in I_n\mapsto\xi(\lambda,n)\in(0,+\infty)$) et $\xi^- = -\xi^+$.

THÉORÈME 3.2 (Représentation spectrale de A). *Les fonctions $\psi^j(\lambda,n,x,z)$, $j=1$ ou 2, $\lambda\in I_n=(\lambda(0,n),+\infty)$, $n\in\mathbb{N}^*$, définies sur Ω par*

$$\varepsilon_j = (-1)^j \quad et \quad \psi^j(\lambda,n,x,z) = (\partial_\lambda\xi(\lambda,n))^{1/2}\,\Psi(\varepsilon_j\xi(\lambda,n),n,x,z) \tag{3.21}$$

$$= (\partial_\lambda\xi(\lambda,n))^{1/2}\,(2\pi)^{-1/2}e^{i\varepsilon_j\xi(\lambda,n)x}\,V(\varepsilon_j\xi(\lambda,n),n,z).$$

sont des fonctions propres généralisées de A. Elles satisfont l'équation différentielle $A\psi^j(\lambda,n,.,.) = \lambda\psi^j(\lambda,n,.,.)$. Si f est une fonction de $L^2(\Omega)$, les coefficients

$$f^j(\lambda,n) = L^2(I_n\,;d\lambda) - \lim_{X\to+\infty}\int_{\Omega\cap\{x/|x|<X\}} f(x,z)\overline{\psi^j(\lambda,n,x,z)}dxdz \tag{3.22}$$

vérifient

$$(f^1,f^2)\in\bigoplus_{n\geq 1}\widetilde{\mathcal{H}}_n,\ \text{avec}\ \widetilde{\mathcal{H}}_n = L^2(I_n)\oplus L^2(I_n). \tag{3.23}$$

La transformation $\widetilde{\mathcal{F}}$ $[f\in L^2(\Omega)\longmapsto (f^1,f^2)\in\oplus\widetilde{\mathcal{H}}_n)]$ est unitaire et on a

$$f(x,z) = \sum_{n\geq 1}\sum_{j=1,2}\int_{I_n} f^j(\lambda,n)\psi^j(\lambda,n,x,z)d\lambda, \tag{3.24a}$$

$$(f/g)_{L^2(\Omega)} = \sum_{n\geq 1}\sum_{j=1,2}\int_{I_n} f^j(\lambda,n)\overline{g^j(\lambda,n)}d\lambda. \tag{3.24b}$$

Si f appartient au domaine $D(A)$, on a, pour $j=1$ ou 2, $\lambda\in I_n$ et $n\in\mathbb{N}^*$,

$$(Af)^j(\lambda,n) = \lambda f^j(\lambda,n). \tag{3.25}$$

Pour $\zeta\notin\sigma(A) = \overline{I}_1 = [\lambda(0,1),+\infty)$, pour f et $g\in L^2(\Omega)$, la résolvante vérifie

$$(R_A(\zeta)f/g)_{L^2(\Omega)} = \sum_{n\geq 1}\sum_{j=1,2}\int_{I_n}\frac{f^j(\lambda,n)\overline{g^j(\lambda,n)}}{\lambda-\zeta}d\lambda. \tag{3.26}$$

Notons que lorsque l'intégrale $\int_\Omega f(x,z)\overline{\psi^j(\lambda,n,x,z)}dxdz$ converge au sens ordinaire, les formules (3.22), (3.21) et (3.14) donnent

$$f^j(\lambda,n) = (\partial_\lambda\xi(\lambda,n))^{1/2}\,\widetilde{f}(\varepsilon_j\xi(\lambda,n),n),\qquad \lambda\in I_n. \tag{3.27}$$

3.2. Principe d'absorption limite. Il s'agit d'étudier la limite de (3.19) ou (3.26) quand ζ tend vers $\mu\in\sigma(A)$. Pour $n\in\mathbb{N}^*$, $\zeta\notin\sigma(A)$, f et $g\in L^2(\Omega)$, nous posons

$$\begin{cases}\widetilde{B}_n(\zeta,f,g) = \displaystyle\int_{\mathbb{R}}\frac{\widetilde{f}(\xi,n)\overline{\widetilde{g}(\xi,n)}}{\lambda(\xi,n)-\zeta}d\xi = \sum_{j=1,2}B_n^j(\zeta,f,g)\\[2ex] \text{avec } B_n^j(\zeta,f,g) = \displaystyle\int_{I_n}\frac{f^j(\lambda,n)\overline{g^j(\lambda,n)}}{\lambda-\zeta}d\lambda.\end{cases} \tag{3.28}$$

Nous introduisons, comme dans §2.2.2, les espaces à poids $L^2_s(\Omega)$, s réel, et nous pouvons définir des opérateurs de trace associés à la transformation unitaire du théorème 3.1.

PROPOSITION 3.2. *Soit $s>1/2$. Pour n entier strictement positif et ξ réel, on peut définir $\widetilde{\tau}_n(\xi)$ forme linéaire continue sur $L^2_s(\Omega)$ telle que*

$$\begin{cases} \text{si } f\in C_0^\infty(\Omega)\ ,\ \widetilde{\tau}_n(\xi)f = \widetilde{f}(\xi,n) = \displaystyle\int_\Omega f(x,z)\overline{\Psi(\xi,n,x,z)}dxdz,\\ \text{si } f\in L^2_s(\Omega)\ ,\ \widetilde{\tau}_n(\xi)f = (\widehat{f}(\xi,.)\,/V(\xi,n,.))_{L^2((0,H))}.\end{cases} \tag{3.29}$$

La fonction $\xi \longmapsto \widetilde{\tau}_n(\xi)$ *définie sur* $\mathbb{R}$ *à valeurs dans* $B(L^2_s(\Omega),\mathbb{C})$ *est localement höldérienne d'exposant* $\delta \in [0,1]$, $\delta < s - 1/2$. *Il existe* $M(\xi,\xi') = M_n(s,\delta,\xi,\xi')$ *fonction continue en* (ξ,ξ') *telle que*

$$\forall f \in L^2_s(\Omega)\ ,\ |\widetilde{\tau}_n(\xi)f - \widetilde{\tau}_n(\xi')f| \leq M(\xi,\xi')|\xi - \xi'|^\delta \|f\|_{L^2_s(\Omega)}. \tag{3.30}$$

Preuve. Si $s > 1/2$ et si $f \in L^2_s(\Omega)$, pour presque tout $z \in (0,H)$, on a $f(.,z) \in L^2_s(\mathbb{R})$ et la trace en $\xi \in \mathbb{R}$ de la transformée de Fourier $\widehat{f}(.,z)$ est définie. De plus

$$|\widehat{f}(\xi,z)| \leq (2\pi)^{-1/2}\left(\int_{-\infty}^{+\infty}(1+x^2)^{-s}dx\right)^{1/2}\left(\int_{-\infty}^{+\infty}|f(x,z)|^2(1+x^2)^s dx\right)^{1/2}.$$

On en déduit que $\widehat{f}(\xi,.)$ appartient à $L^2((0,H))$, ainsi que l'estimation de continuité

$$|\ (\widehat{f}(\xi,.)\ /\ V(\xi,n,.))_{L^2((0,H))}\ | \leq C(s)\|f\|_{L^2_s(\Omega)}. \tag{3.31}$$

Il suffit d'établir (3.30) pour $f \in C_0^\infty(\Omega)$. On a

$$\begin{aligned}
&(2\pi)^{1/2}|\ \widetilde{f}(\xi,n) - \widetilde{f}(\xi',n)\ | \\
&= \left| \int_\Omega f(x,z)e^{-i\xi x}\Big(V(\xi,n,z) - V(\xi',n,z)\Big)\,dxdz \right. \\
&\qquad \left. + \int_\Omega f(x,z)(e^{-i\xi x} - e^{-i\xi' x})V(\xi',n,z)\,dxdz \right| \\
&\leq \int_{-\infty}^{+\infty}\left[\int_0^H |f(x,z)|^2\,dz\right]^{1/2}\|V(\xi,n,.) - V(\xi',n,.)\|_{L^2((0,H))}\,dx \\
&\qquad + \int_{-\infty}^{+\infty}\left[\int_0^H |f(x,z)|^2\,dz\right]^{1/2}|e^{-i\xi x} - e^{-i\xi' x}|\,dx.
\end{aligned}$$

Le théorème 4.5 de l'annexe donne, avec $C_n(\xi,\xi')$ continue en (ξ,ξ'),

$$\|V(\xi,n,.) - V(\xi',n,.)\|_{L^2((0,H))} \leq C_n(\xi,\xi')|\xi - \xi'|.$$

La fonction exponentielle satisfait, avec $\delta \in [0,1]$,

$$|e^{-i\xi x} - e^{-i\xi' x}| \leq 2|\xi - \xi'|^\delta |x|^\delta.$$

On obtient alors (3.30) sous les conditions $s > 1/2$ et $\delta < s - 1/2$. □

L'estimation (3.31) de continuité sur $L^2_s(\Omega)$ pour les formes $\widetilde{\tau}_n(\xi)$, ξ réel, s'écrit encore

$$\forall f \in L^2_s(\Omega),\ |\widetilde{\tau}_n(\xi)f| \leq C\|f\|_{L^2_s(\Omega)} \ \text{ avec } C = C(s). \tag{3.32}$$

Remarquons que la forme $\widetilde{\tau}_n(0)$ est non nulle. Par suite, l'estimation de type (2.54) avec $\delta > 0$ n'est vraie que sur le noyau de $\widetilde{\tau}_n(0)$. A partir de (3.30), elle s'écrit

$$\forall f \in \ker \widetilde{\tau}_n(0),\ |\widetilde{\tau}_n(\xi)f| \leq M(\xi,0)|\xi|^\delta \|f\|_{L^2_s(\Omega)}. \tag{3.33}$$

Donnons quelques propriétés de ce noyau.

PROPOSITION 3.3. *Pour $n \geq 1$ et $s > 1/2$, on pose*

$$NL^2_s(n) = ker\, \widetilde{\tau}_n(0) = \{f \in L^2_s(\Omega)/(\widehat{f}(0,.)/V(0,n,.))_{L^2((0,H))} = 0\}, \tag{3.34a}$$

$$NL^2_s = \bigcap_{n\geq 1} NL^2_s(n) = \{f \in L^2_s(\Omega) \,/\, \widehat{f}(0,.) = 0\}. \tag{3.34b}$$

L'intersection de ces espaces avec $C^\infty_0(\Omega)$ est indépendante de s, et décrite par

$$W_n = NL^2_s(n) \cap C^\infty_0(\Omega) = \left\{ f \in C^\infty_0(\Omega) \,/\, \int_\Omega f(x,z)V(0,n,z)dxdz = 0 \right\}, \tag{3.35a}$$

$$W = NL^2_s \cap C^\infty_0(\Omega) = \left\{ f \in C^\infty_0(\Omega) \,/\, \int_{I\!R} f(x,z)dx = 0 \right\}. \tag{3.35b}$$

Chacun des espaces W_n ou W est dense dans $L^2_s(\Omega)$ si $s \leq 1/2$. L'espace W_n (respectivement W) est dense dans $NL^2_s(n)$ (respectivement NL^2_s) si $s > 1/2$.

Preuve. Lorsque $f \in L^2_s(\Omega)$, on a

$$\widetilde{\tau}_n(0)f = \widetilde{f}(0,n) = (\widehat{f}(0,.) \,/\, V(0,n,.))_{L^2((0,H))},$$

et la description de NL^2_s, W_n ou W en résulte (on rappelle que les fonctions $V(0,n,.)$, $n \geq 1$, forment une base orthonormale de $L^2((0,H))$).

La densité dans $L^2_s(\Omega)$ de W, et donc celle de W_n, découle, pour $s \leq 1/2$, de la densité de $C^\infty_0(I\!R) \otimes C^\infty_0((0,H))$ dans $L^2_s(\Omega)$, et de la densité de l'espace

$$\left\{ f \in C^\infty_0(I\!R) \,/\, \int_{-\infty}^{+\infty} f(x)dx = 0 \right\} \text{ dans } L^2_s(I\!R) \text{ (résultat en défaut si } s > 1/2).$$

Justifions enfin le dernier résultat de densité. Soient $s > 1/2$, $f \in NL^2_s(n)$ (respectivement NL^2_s) et $\varepsilon > 0$. Il existe $f_\varepsilon \in C^\infty_0(\Omega)$ telle que $\|f - f_\varepsilon\|_{L^2_s(\Omega)} < \varepsilon$. Considérons

$$F_\varepsilon(x,z) = f_\varepsilon(x,z) - (\widetilde{\tau}_n(0)f_\varepsilon)\varphi(x)\psi_n(z)$$

(respectivement $F_\varepsilon(x,z) = f_\varepsilon(x,z) - \varphi(x)\widehat{f}_\varepsilon(0,z)$), où φ et ψ_n sont choisies telles que

$$\varphi \in C^\infty_0(I\!R), \int_{-\infty}^{+\infty} \varphi(x)dx = 1,$$

$$\psi_n \in C^\infty_0((0,H)), \int_0^H \psi_n(z)V(0,n,z)dz = 1.$$

Cette fonction appartient à W_n (respectivement W) et satisfait, grâce à l'appartenance de f à $NL^2_s(n)$ (respectivement NL^2_s), $\lim_{\varepsilon\to 0} \|f - F_\varepsilon\|_{L^2_s(\Omega)} = 0$. □

On définit aussi des opérateurs de trace pour la représentation spectrale du théorème 3.2; ils seront utiles pour étudier le comportement de $\widetilde{B}_n(\zeta, f, g)$ au voisinage des seuils $\lambda(0,n)$. On montre au théorème 4.5 de l'annexe que ces seuils sont des zéros d'ordre exactement 2 pour les fonctions $\xi \mapsto \lambda(\xi,n) - \lambda(0,n)$.

PROPOSITION 3.4. *On se donne le réel $s > 1/2$ et les entiers $n \geq 1$, $j = 1$ ou 2. Pour tout réel $\lambda \geq \lambda(0,n)$, on peut définir $\tau_n^j(\lambda)$, forme linéaire continue sur $L_s^2(\Omega)$ par*

$$\text{(3.36a)} \quad \varepsilon_j = (-1)^j \quad et \quad \tau_n^j(\lambda)f = (\partial_\lambda \xi(\lambda,n))^{1/2} \widetilde{\tau}_n(\varepsilon_j \xi(\lambda,n))f \ si \ \lambda > \lambda(0,n),$$

$$\text{(3.36b)} \qquad \tau_n^j(\lambda(0,n))f = 0.$$

La limite quand λ tend vers $\lambda(0,n)_+$ de $\tau_n^j(\lambda)f$, avec f fonction quelconque de l'espace $L_s^2(\Omega)$, peut être infinie. On a l'existence de $C(\lambda) = C_n(s,\lambda)$, fonction continue par rapport à λ, telle que pour $\lambda > \lambda(0,n)$ et pour $f \in L_s^2(\Omega)$,

$$\text{(3.37)} \quad \tau_n^j(\lambda)f = C(\lambda,f)(\lambda - \lambda(0,n))^{-1/4} \quad et \quad |\, C(\lambda,f)| \leq C(\lambda)\|f\|_{L_s^2(\Omega)}.$$

L'application $\lambda \mapsto \tau_n^j(\lambda)$ est localement höldérienne d'exposant δ. Précisément

(a) *Sur $I_n = (\lambda(0,n), +\infty)$, elle est à valeurs dans $B(L_s^2(\Omega), \mathbb{C})$. Si $\delta \in [0,1]$ et $\delta < s - 1/2$, il existe $M(\lambda,\lambda') = M_n^j(s,\delta,\lambda,\lambda')$ continue en (λ,λ') telle que pour*

$$\text{(3.38)} \qquad f \in L_s^2(\Omega), \ \lambda \ et \ \lambda' > \lambda(0,n)$$

on a

$$\text{(3.39)} \qquad |\tau_n^j(\lambda)f - \tau_n^j(\lambda')f| \leq M(\lambda,\lambda')|\lambda - \lambda'|^\delta \|f\|_{L_s^2(\Omega)}$$

(b) *Sur $\overline{I}_n = [\lambda(0,n), +\infty)$, elle est à valeurs dans $B(NL_s^2(n), \mathbb{C})$ (rappellons que $NL_s^2(n)$ est le noyau de la forme $\widetilde{\tau}_n(0)$, étudié à la proposition 3.3). Sous la condition*

$$\text{(3.40)} \qquad s > 1, \ \delta \in [0, 1/4] \ et \ \delta < (s-1)/2,$$

l'estimation höldérienne (3.39) *est valide pour*

$$\text{(3.41)} \qquad f \in NL_s^2(n), \ \lambda \ et \ \lambda' \geq \lambda(0,n),$$

Preuve. Pour $s > 1/2$, λ et $\lambda' \in I_n$, l'estimation (3.39) sur $L_s^2(\Omega)$ découle directement de la définition (3.36) de $\tau_n^j(\lambda)$, de la proposition 3.2 et de l'analyticité de la fonction $(\xi^+ : \ \lambda \in I_n \mapsto \xi(\lambda,n) \in (0,+\infty))$.

L'étude au voisinage de $\lambda = \lambda(0,n)$ s'appuie sur le développement de la fonction $\xi \mapsto \lambda(\xi,n)$ au voisinage de 0. Selon le théorème 4.5, on a

$$\lambda(\xi,n) - \lambda(0,n) = \xi^2 a_n(\xi) \quad \text{et} \quad \partial_\xi \lambda(\xi,n) = \xi \, b_n(\xi),$$

d'où l'on tire

$$\xi = (\lambda(\xi,n) - \lambda(0,n))^{1/2} \, G_n(\xi),$$
$$\partial_\lambda \xi(\lambda,n) = (\lambda(\xi,n) - \lambda(0,n))^{-1/2} \, H_n(\xi),$$
$$\tau_n^j(\lambda)f = (\lambda(\xi,n) - \lambda(0,n))^{-1/4} \, H_n(\xi)^{1/2} \, \widetilde{\tau}_n(\varepsilon_j \xi)f.$$

Dans ces égalités, les fonctions a_n, b_n, G_n et H_n sont analytiques et ne s'annulent pas sur $\mathbb{R}$.

L'estimation (3.32) donne alors (3.37).

L'estimation (3.33) donne, si $\lambda \in I_n$, si $f \in NL^2_s(n)$, si $\delta' \in [0,1]$ et $\delta' < s - 1/2$, la majoration:

$$|\tau_n^j(\lambda) f| \leq (\lambda - \lambda(0,n))^{(-1+2\delta')/4}\, M(\lambda) \|f\|_{L^2_s(\Omega)}.$$

L'estimation höldérienne (3.39) au voisinage de $\lambda(0,n)$ est donc obtenue si $\delta' > 1/2$, ce qui nécessite $s > 1$. L'exposant δ est alors majoré par $1/4$ puisque $\delta' \leq 1$, et par $(s-1)/2$ puisque $\delta' < s - 1/2$. □

On peut établir maintenant des estimations, uniformes par rapport à ζ, pour les termes (3.28) intervenant dans la résolvante, comme il a été fait dans §2.2.3.

PROPOSITION 3.5. *On se donne les deux réels $s \geq 0$ et $\Lambda > 0$, et on définit l'entier*

$$N(\Lambda) = \mathrm{Min}\{n \in \mathbb{N}^* \,/\, \Lambda < \lambda(0,n)\}. \tag{3.42}$$

Alors il existe une constante $C = C(\Lambda, s)$ telle que

$$\forall f, g \in L^2_s(\Omega), \sum_{n \geq N(\Lambda)} |\widetilde{B}_n(\zeta, f, g)| \leq C \|f\|_{L^2_s(\Omega)}\, \|g\|_{L^2_s(\Omega)} \tag{3.43}$$

lorsque le nombre complexe ζ satisfait

$$\zeta \notin \sigma(A) \quad et \quad |\zeta| \leq \Lambda. \tag{3.44}$$

PROPOSITION 3.6. *On se donne l'entier $n \geq 1$ et le réel $\Lambda > 1$ tels que $\lambda(0,n) \leq \Lambda$. Alors il existe une constante $C = C(\Lambda, s)$ telle que*

$$|\widetilde{B}_n(\zeta, f, g)| \leq C \|f\|_{L^2_s(\Omega)}\, \|g\|_{L^2_s(\Omega)} \tag{3.45}$$

lorsque le réel s, les fonctions f et g, et le nombre complexe ζ satisfont

(a) $s > 1/2$, f *et* $g \in L^2_s(\Omega)$, $|\zeta - \lambda(0,n)| \geq \Lambda^{-1}$ *et* (3.44);

(b) $s > 1$, f *et* $g \in NL^2_s(n)$, (3.44).

Preuve. Elle est à rapprocher de celle de la proposition 2.7.

Pour ζ satisfaisant (3.44), on note $\lambda' = \mathrm{Re}\zeta$ et on considère les zéros de $\xi \mapsto \lambda(\xi, n) - \lambda'$ qui sont simples et égaux à $\pm\xi = \pm\xi(\lambda', n)$ si $\lambda' \neq \lambda(0,n)$, et qui se réduisent à $\xi = 0$, zéro d'ordre 2, si $\lambda' = \lambda(0,n)$.

Lorsque $|\zeta - \lambda(0,n)|$ est minoré, l'estimation (3.30) pour les opérateurs $\widetilde{\tau}_n(\xi)$ conduit à (3.45) si on s'impose de plus la condition (3.44) pour ζ et si on choisit $s > 1/2$.

Lorsque ζ peut s'approcher du seuil $\lambda(0,n)$, on ramène l'estimation de $\widetilde{B}_n(\zeta, f, g)$ à celle de

$$c^j_\alpha(\zeta, f, g) = \int_{\lambda'-\alpha}^{\lambda'+\alpha} \frac{f^j(\lambda, n)\overline{g^j(\lambda, n)}}{\lambda - \lambda' - i\mathrm{Im}\zeta} d\lambda,$$

où $\alpha > 0$ est fixé. L'estimation (3.39) pour l'opérateur $\tau_n^j(\lambda)$, sous les conditions (3.40) et (3.41), conduit à (3.45) avec $s > 1$, f et g dans $\ker \widetilde{\tau}_n(0) = NL^2_s(n)$ et ζ satisfaisant (3.44).

On remarquera que l'estimation de c^j_α est un peu plus simple que celle faite dans la proposition 2.7 car le dénominateur de la fonction à intégrer est de degré 1 en λ et on utilise (3.39) avec $\delta > 0$. □

On ne peut pas reprendre telle quelle la démarche d'Agmon pour obtenir le principe d'absorption limite fort au voisinage des seuils $\lambda(0,n)$, comme nous l'avons fait dans le cas d'une stratification verticale au §2. En effet, dans un tel voisinage de $\lambda(0,n)$, on obtient facilement, à partir des propositions 3.5 et 3.6, un principe d'absorption limite faible dans le dual E' de l'espace $E = NL^2_s(n)$, c'est-à-dire l'existence de la limite faible dans E', pour f appartenant à E et quand $\pm\mathrm{Im}\zeta$ tend vers 0_+, de $R_A(\zeta)f$. A ce point-là, on ne peut plus agir de même pour transformer cette limite faible en limite forte. De plus, la démarche d'Agmon ne donne pas directement les propriétés höldériennes de la résolvante.

Nous allons donc commencer par établir des propriétés höldériennes pour les fonctions $\widetilde{B}_n(.,f,g)$ sur $\mathbb{C}\backslash\sigma(A)$, à partir desquelles nous obtiendrons immédiatement le principe d'absorption limite fort.

PROPOSITION 3.7. *On se donne l' entier $n_0 \geq 1$, et les réels Λ_1 et Λ_2 tels que*

$$\lambda(0,n_0-1) < \Lambda_1 < \lambda(0,n_0) < \Lambda_2 < \lambda(0,n_0+1). \tag{3.46}$$

Alors, pour n entier strictement positif, il existe une constante $C_n = C_n(\Lambda_1,\Lambda_2,s)$ telle que

$$|\widetilde{B}_n(\zeta,f,g) - \widetilde{B}_n(\zeta',f,g)| \leq C_n|\zeta-\zeta'|^\delta \|f\|_{L^2_s(\Omega)}\ \|g\|_{L^2_s(\Omega)} \tag{3.47}$$

d'une part lorsque les nombres complexes ζ et ζ' satisfont

$$\zeta \text{ et } \zeta' \notin \sigma(A),\ |\zeta| \text{ et } |\zeta'| \in (\Lambda_1,\Lambda_2), \tag{3.48}$$

et d'autre part sous les conditions suivantes pour l'entier n et pour les parties imaginaires de ζ et ζ' , pour les réels s et δ, pour les fonctions f et g,

(a) *Si n est différent de n_0,*

$$\begin{cases} s > 1/2 \ \text{ et } \ f,g \in L^2_s(\Omega), \\ \text{si } n \geq n_0+1, \text{ alors } \delta \in [0,1] \quad \left(\text{et } \displaystyle\sum_{n\geq n_0+1} C_n < +\infty\right), \\ \text{si } n \leq n_0-1 \ \text{ et } \mathrm{Im}\,\zeta \,.\, \mathrm{Im}\,\zeta' > 0, \text{ alors } \delta \in [0,\mathrm{Min}(1,s-1/2)). \end{cases} \tag{3.49}$$

(b) *Si n est égal à n_0 et* $\mathrm{Im}\ \zeta.\ \mathrm{Im}\ \zeta' > 0$,

$$\begin{cases} s > 1 \ \text{ et } \ f,g \in NL^2_s(n_0), \\ \delta \in [0,\mathrm{Min}(1/4,(s-1)/2)). \end{cases} \tag{3.50}$$

Preuve. On a pour $n \geq 1$, ζ et $\zeta' \notin \sigma(A)$, f et $g \in L^2_s(\Omega)$:

$$\widetilde{B}_n(\zeta,f,g) - \widetilde{B}_n(\zeta',f,g) = \sum_{j=1}^{2}(\zeta-\zeta')\int_{I_n} \frac{f^j(\lambda,n)\overline{g^j(\lambda,n)}}{(\lambda-\zeta)(\lambda-\zeta')}d\lambda.$$

On décompose l'intervalle d'intégration $I_n = (\lambda(0,n),+\infty)$, lorsque l'indice n est inférieur ou égal à n_0, de la manière suivante:

$I_n = (\lambda(0,n),\Lambda_1] \cup (\Lambda_1,\Lambda_2) \cup [\Lambda_2,+\infty)$ si $n \leq n_0-1$,

$I_{n_0} = (\lambda(0,n_0),\Lambda_2) \cup [\Lambda_2,+\infty)$.

Sur chaque intervalle, on mène une estimation de l'intégrale, qui peut être rapprochée de celle des propositions 3.5 ou 3.6, bien qu'elle soit plus délicate, en utilisant la propriété höldérienne (3.39) des traces $\tau_n^j(\lambda)f = f^j(\lambda, n)$. □

On peut maintenant donner des propriétés höldériennes de la résolvante sur $\mathbb{C}\backslash\sigma(A)$.

THÉORÈME 3.3. *On se donne le réel $\Lambda > 1$ et l'entier $N = N(\Lambda)$ associé par (3.42) et caractérisé par $\lambda(0, N-1) \leq \Lambda < \lambda(0, N)$. On choisit le réel $\widetilde{\Lambda}$ strictement positif tel que $\widetilde{\Lambda} < \Lambda' = \Lambda'(\Lambda) = \mathrm{Min}\{\lambda(0, n+1) - \lambda(0, n) \ / \ 1 \leq n < N\}$. On se donne des réels s et δ, une fonction f et des nombres complexes ζ et ζ' dans $\mathbb{C}^{\pm} = \{\zeta \in \mathbb{C} \ / \ \pm \mathrm{Im}\,\zeta \geq 0\}$.*

Alors sous les conditions suivantes portant sur s, sur f, sur Z égal à ζ ou ζ', et sur δ,

$$\begin{cases} s > 1/2 \ \text{ et } f \in E_s = L^2_s(\Omega) \ \text{ (donc } E'_s = L^2_{-s}(\Omega)), \\ Z \in \mathbb{C}^{\pm},\ Z \notin \sigma(A),\ |Z| \leq \Lambda, \ \text{ et } |Z - \lambda(0,n)| \geq \widetilde{\Lambda} \ \text{ pour tout entier } n < N, \\ \delta \in [0, \mathrm{Min}(1, s - 1/2)), \end{cases} \tag{3.51}$$

ou sous les conditions suivantes portant sur s, sur n entier fixé, sur f, sur Z égal à ζ ou ζ', et sur δ,

$$\begin{cases} s > 1,\ n < N, \ \text{ et } f \in E_s = NL^2_s(n), \\ Z \in \mathbb{C}^{\pm},\ Z \notin \sigma(A), \ \text{ et } |Z - \lambda(0,n)| \leq \widetilde{\Lambda}, \\ \delta \in [0, \mathrm{Min}(1/4, (s-1)/2), \end{cases} \tag{3.52}$$

il existe une constante $C = C(\Lambda, s, \delta)$ telle que

$$\forall f \in E_s, \|R_A(\zeta)f - R_A(\zeta')f\|_{E'_s} \leq C|\zeta - \zeta'|^{\delta} \ \|f\|_{L^2_s(\Omega)}, \tag{3.53}$$

$$\forall f \in E_s, \ \|R_A(\zeta)f\|_{E'_s} \leq C\|f\|_{L^2_s(\Omega)}. \tag{3.54}$$

L'estimation au voisinage de $\lambda(0,n)$ est optimale au sens où il existe f_n dans $L^2_s(\Omega)$, f_n n'appartenant pas à l'hyperplan fermé $NL^2_s(n)$, telle que

$$\lim_{\zeta \to \lambda(0,n),\ \zeta \notin \sigma(A)} \|R_A(\zeta)f_n\|_{L^2_{-s}(\Omega)} = +\infty. \tag{3.55}$$

Preuve. Sous les conditions (3.51) ou (3.52), la proposition 3.7 donne (3.53), les propositions 3.5 et 3.6 donnent (3.54).

Pour réaliser (3.55), on choisit

$$f_n(x,z) = (2\pi)^{-1/2} \int_{-\infty}^{+\infty} \rho(\xi) e^{i\xi x} V(\xi, n, z) d\xi,$$

avec ρ fonction fixée dans $C_0^{\infty}(\mathbb{R})$ telle que $\rho(\xi) = 1$ si $|\xi| \leq 1$.

(a) On a

$$f_n \in \bigcap_{s \geq 0} L^2_s(\Omega) = \bigcap_{k \in \mathbb{N}} L^2_k(\Omega).$$

En effet, f_n est la transformée de Fourier de $\rho V(.,n,z)$ et on a l'estimation

$$\|x^k f_n(x,z)\|_{L^2(\Omega;dxdz)} \leq C(k)\| \sum_{k'=0}^{k} \rho^{(k')}(\xi)\partial_\xi^{(k-k')} V(\xi,n,z)\|_{L^2(\Omega;d\xi dz)} \leq C(k,\rho,n),$$

car ρ est à support compact et la fonction $\xi \mapsto V(\xi,n,.)$ est de classe C^∞ sur $\mathbb{R}$ à valeurs dans $L^2((0,H))$ (cf. le théorème 4.5 de l'annexe).

(b) On a

$$(R_A(\lambda(0,n)+i\varepsilon)f_n, f_n) = \int_{-\infty}^{+\infty} \frac{|\rho(\xi)|^2}{\lambda(\xi,n)-\lambda(0,n)-i\varepsilon} d\xi = I_1(\varepsilon) + I_2(\varepsilon)$$

avec

$$|I_1(\varepsilon)| = |\int_{|\xi|\geq 1} \frac{|\rho(\xi)|^2}{\lambda(\xi,n)-\lambda(0,n)-i\varepsilon} d\xi| \leq (\lambda(1,n)-\lambda(0,n))^{-1}\|\rho\|^2_{L^2(\mathbb{R})},$$

et

$$|I_2(\varepsilon)| = |\int_{-1}^{1} \frac{1}{\lambda(\xi,n)-\lambda(0,n)-i\varepsilon} d\xi|$$

qui tend vers $+\infty$ lorsque ε tend vers 0. En effet, l'encadrement $\xi^2 c_m^2 \leq \lambda(\xi,n) - \lambda(0,n) = \xi^2 a_n(\xi) \leq \xi^2 \|c\|^2_{L^\infty((0,H))}$ établi au théorème 4.5, permet l'estimation

$$|I_2(\varepsilon)| \geq \int_{-1}^{1} \frac{\xi^2 a_n(\xi)}{\xi^4 a_n^2(\xi)+\varepsilon^2} d\xi \geq \frac{2m_1}{M_1} \int_0^1 \frac{\xi^2}{\xi^3+\varepsilon^2 M_1^{-1}} d\xi \geq \frac{2m_1}{3M_1} |\mathrm{Log}\, \varepsilon^2 M_1^{-1}|,$$

avec

$$m_1 = \operatorname*{Min}_{0\leq\xi\leq 1} a_n(\xi) > 0 \quad \text{et} \quad M_1 = \operatorname*{Max}_{0\leq\xi\leq 1} a_n^2(\xi) > 0. \qquad \square$$

COROLLAIRE 3.2. *Sous les conditions* (3.51), *on a*

$$C_1\|\nabla R_A(\zeta)f\|_{(L^2_{-s}(\Omega))^2} \leq \|R_A(\zeta)f\|_{L^2_{-s}(\Omega)} + \|AR_A(\zeta)f\|_{L^2_{-s}(\Omega)} \leq C_2\|f\|_{L^2_s(\Omega)}. \tag{3.56}$$

Sous les conditions (3.52), on peut aussi évaluer $\nabla R_A(\zeta)f$. Nous y reviendrons dans l'article suivant (partie II).

La complétude de l'espace $B(E,E')$, avec E espace de Banach, pour la topologie de la norme des opérateurs bornés, donne alors directement le principe d'absorption limite.

THÉORÈME 3.4 (Principe d'absorption limite). *On se donne les réels* μ, s *et l'espace* $E_s(\mu)$ *tels que*

$$\begin{cases} \mu \in \sigma(A) \setminus \{\lambda(0,n)/n \in \mathbb{N}^*\},\ s > 1/2 \ \text{ et } E_s(\mu) = L^2_s(\Omega), \\ \quad ou \\ \mu = \lambda(0,n) \ \text{ avec } n \in \mathbb{N}^*,\ s > 1 \ \text{ et } E_s(\mu) = NL^2_s(n). \end{cases} \tag{3.57}$$

(i) *Les deux limites suivantes existent pour la topologie de la norme dans* $B(E_s(\mu), E_s(\mu)')$:

$$R_A^{\pm}(\mu) = \lim_{\zeta\to\mu,\ \pm\mathrm{Im}\zeta>0} R_A(\zeta), \tag{3.58}$$

(ii) *Pour f fonction donnée dans $E_s(\mu)$, $u^{\pm} = R_A^{\pm}(\mu)f$ vérifie par dualité:*

$$(A - \mu I)u^{\pm} = f. \tag{3.59}$$

Remarque 3.1. Les fonctions $\zeta \mapsto R_A^{\pm}(\zeta)$ ainsi définies sur $\mathbb{C}^{\pm} = \{\zeta \in \mathbb{C} \ / \ \pm \mathrm{Im}\zeta \geq 0\}$ sont localement höldériennes,

(a) d'exposant $\delta \in [0, \mathrm{Min}(1, s - 1/2))$ dans l'ouvert $\mathbb{C}^{\pm} \backslash S$, où S est l'ensemble des seuils $\lambda(0,n), n \geq 1$, et à valeurs dans $B(L_s^2(\Omega), L_{-s}^2(\Omega))$,

(b) d'exposant $\delta \in [0, \mathrm{Min}(1/4, (s-1)/2))$ dans un voisinage V_n de $\lambda(0,n)$ tel que $V_n \cap S = \{\lambda(0,n)\}$, et à valeurs dans $B(NL_s^2(n), NL_s^2(n)')$.

Remarque 3.2. L'espace $NL_s^2(n)'$ n'est pas un espace de distributions bien que tout élément de $NL_s^2(n)'$ admette des prolongements distributions qui sont dans l'espace $L_{-s}^2(\Omega)$. Dans l'article suivant, nous préciserons les propriétés de la résolvante aux seuils.

Exprimons maintenant les valeurs limites de la résolvante, à l'aide de crochets de dualité.

PROPOSITION 3.8. *Soit $\mu \in \sigma(A)$. On se donne s réel et f, g fonctions dans l'espace E_s tels que*

$$s > 1/2 \quad et \quad E_s = L_s^2(\Omega) \quad si\ \mu \in (\lambda(0, m-1), \lambda(0,m)), \tag{3.60}$$

$$s > 1 \quad et \quad E_s = NL_s^2(m) \quad si\ \mu = \lambda(0,m). \tag{3.61}$$

Alors on a

$$< R_A^{\pm}(\mu)f, g >_{E_s', E_s} = \begin{cases} \displaystyle\sum_{n \geq m} \int_{\mathbb{R}} \frac{\widetilde{f}(\xi,n)\overline{\widetilde{\overline{g}}(\xi,n)}}{\lambda(\xi,n) - \mu} d\xi \\ \displaystyle + \sum_{n<m} v.\,p. \int_{\mathbb{R}} \frac{\widetilde{f}(\xi,n)\overline{\widetilde{\overline{g}}(\xi,n)}}{\lambda(\xi,n) - \mu} d\xi \\ \displaystyle \pm i\pi \sum_{n<m} \sum_{j=1}^{2} f^j(\mu,n)\overline{\overline{g}^j(\mu,n)}. \end{cases} \tag{3.62}$$

Remarque 3.3. Le cas particulier $c_1 = c_2 = c$ relève de ce théorème. L'hypothèse **(H)** est satisfaite. Les fonctions propres de l'opérateur B_ξ peuvent être choisies indépendantes de ξ. Une base de telles fonctions est donnée par la suite des fonctions (2.9): $V(\xi, n, z) = V_n(z)$, $n \in \mathbb{N}^*$. Les estimations fondamentales pour $\widetilde{B}_n(\zeta, f, g)$ au voisinage des seuils (cf. les propositions 3.6 et 3.7) sont valides avec le réel $s > 1$, et avec les fonctions f et g dans l'espace

$$NL_s^2(n) = \left\{ f \in L_s^2(\Omega) \ / \ \int_\Omega f(x,z)\, V_n(z) dx dz = 0 \right\}.$$

Ces résultats sont à rapprocher de ceux de [W87]. Le choix $f(x,z) = g(x,z) = \widehat{\rho}(x)V_n(z)$, avec ρ choisie comme dans la preuve du théorème 3.3, permet de montrer l'optimalité de l'espace $NL_s^2(n)$.

4. Annexe sur les opérateurs de Sturm–Liouville.

4.1. L'opérateur B_n sur $L^2(\mathbb{R})$. On considère un profil $c(x)$ égal à c_1 si $x < 0$, et à c_2 si $x > 0$: cf. (1.5) et la figure 4.1.

Nous choisissons $c_1 > c_2$, quitte à changer l'orientation de l'axe Ox.

L'opérateur

$$B_n = -\frac{d}{dx}\left(c^2(x)\frac{d}{dx}\right) + c^2(x)q_n^2$$

et son domaine $D(B_n) = \left\{u \in H^1(\mathbb{R}) \,/\, c^2u' \in H^1(\mathbb{R})\right\}$ ont été introduits en (2.10).

Suivant la théorie de Weyl–Kodaira, développée dans les livres de Dunford et Schwartz [DS] et de Wilcox [Wi84], nous déterminons $\widetilde{\mathcal{F}}_n$, une représentation spectrale de l'opérateur B_n, selon les étapes E1–E4 qui suivent. Les calculs correspondants sont développés dans le rapport [CD].

Etape E1. Les fonctions propres généralisées de B_n. Ce sont les solutions du problème

$$\text{(4.1)} \qquad \begin{cases} \text{trouver } (\lambda, \Phi) \in \mathbb{R}^+ \times (D(B_n))_{loc}\,, \\ B_n\Phi = \lambda\Phi \text{ sur } \mathbb{R}, \\ \Phi \text{ bornée et non identiquement nulle.} \end{cases}$$

THÉORÈME 4.1. *Les solutions de* (4.1) *existent si et seulement si* $\lambda \in (c_2^2q_n^2, +\infty)$.

(a) *Lorsque* $\lambda \in (c_2^2q_n^2, c_1^2q_n^2]$, *elles sont données par*

$$\text{(4.2)} \qquad \Phi(\lambda, n, x) = a^0(\lambda, n)\Phi^0(\lambda, n, x),$$

où le coefficient $a^0(\lambda, n)$ *est la coordonnée scalaire de* Φ *sur la fonction*

$$\text{(4.3)} \qquad \Phi^0(\lambda, n, x) = \begin{cases} e^{\xi_1' x} & \text{si } x < 0, \\ \cos(\xi_2 x) + \dfrac{c_1^2\xi_1'}{c_2^2\xi_2}\sin(\xi_2 x) & \text{si } x > 0, \end{cases}$$

avec

$$\xi_1' = \xi_1'(\lambda, n) = \frac{1}{c_1}(c_1^2q_n^2 - \lambda)^{1/2} \quad \text{et} \quad \xi_2 = \xi_2(\lambda, n) = \frac{1}{c_2}(\lambda - c_2^2q_n^2)^{1/2}.$$

(b) *Lorsque* $\lambda \in (c_1^2q_n^2, +\infty)$, *elles sont données par*

$$\text{(4.4)} \qquad \Phi(\lambda, n, x) = a^1(\lambda, n)\Phi^1(\lambda, n, x) + a^2(\lambda, n)\Phi^2(\lambda, n, x),$$

où les coefficients $a^i(\lambda, n)$, $i = 1$ *ou* 2, *sont les coordonnées scalaires de* Φ *sur les deux fonctions indépendantes*

$$\text{(4.5)} \qquad \Phi^1(\lambda, n, x) = \begin{cases} \sin(\xi_1 x) & \text{si } x < 0, \\ \dfrac{c_1^2\xi_1}{c_2^2\xi_2}\sin(\xi_2 x) & \text{si } x > 0, \end{cases}$$

$$\text{(4.6)} \qquad \Phi^2(\lambda, n, x) = \begin{cases} \cos(\xi_1 x) & \text{si } x < 0, \\ \cos(\xi_2 x) & \text{si } x > 0, \end{cases}$$

avec $\xi_i = \xi_i(\lambda, n) = \dfrac{1}{c_i}(\lambda - c_i^2q_n^2)^{1/2}$ *pour* $i = 1$ *ou* 2.

En particulier, le spectre ponctuel de B_n est vide et

$$\sigma(B_n) = \sigma_c(B_n) = [c_2^2 q_n^2, +\infty). \tag{4.7}$$

Remarque 4.1. La fonction propre généralisée $\Phi^2(\lambda, n, x)$ est encore définie pour $\lambda = c_1^2 q_n^2$ et on a

$$\Phi^2(c_1^2 q_n^2, n, x) = \Phi^0(c_1^2 q_n^2, n, x) = \begin{cases} 1 \text{ si } x < 0, \\ \cos\left[\, \xi_2(c_1^2 q_n^2, n) x \,\right] \text{ si } x > 0. \end{cases} \tag{4.8}$$

Etape E2. La résolvante $R_{B_n}(\zeta)$ pour ζ non réel. Soit ζ un nombre complexe non réel. On définit

$$\zeta_1 = \frac{1}{c_1}(\zeta - c_1^2 q_n^2)^{1/2}, \qquad \zeta_2 = \frac{1}{c_2}(\zeta - c_2^2 q_n^2)^{1/2}, \tag{4.9}$$

la détermination choisie pour la racine carrée étant celle à partie imaginaire positive ou nulle.

PROPOSITION 4.1. *L'équation $B_n\Phi = \zeta\Phi$ admet pour base de solutions*

$$\Phi^1(\zeta, n, x) = \begin{cases} \sin(\zeta_1 x) & \textit{si } x < 0, \\ \dfrac{c_1^2 \zeta_1}{c_2^2 \zeta_2} \sin(\zeta_2 x) & \textit{si } x > 0, \end{cases}$$

$$\Phi^2(\zeta, n, x) = \begin{cases} \cos(\zeta_1 x) & \textit{si } x < 0, \\ \cos(\zeta_2 x) & \textit{si } x > 0. \end{cases}$$

Ces solutions sont continues en (ζ, x) sur

$$I_n^{\pm} = \{\zeta \in \mathbb{C}/\mathrm{Re}\zeta \in I_n \textit{ et } \pm \mathrm{Im}\zeta \in \overline{\mathbb{R}^+}\} \textit{ avec } I_n = (c_2^2 q_n^2, c_1^2 q_n^2),$$

$$J_n^{\pm} = \{\zeta \in \mathbb{C}/\mathrm{Re}\zeta \in J_n \textit{ et } \pm \mathrm{Im}\zeta \in \overline{\mathbb{R}^+}\} \textit{ avec } J_n = (c_1^2 q_n^2, +\infty).$$

THÉORÈME 4.2. *La résolvante $R_{B_n}(\zeta)f = (B_n - \zeta I)^{-1} f$, pour tout nombre complexe ζ n'appartenant pas au spectre $\sigma(B_n)$ et pour toute fonction f de $C_0^\infty(\mathbb{R})$, est donnée par*

$$R_{B_n}(\zeta)f = \begin{cases} (\mathcal{B}_n(\zeta)f)\, e^{-i\zeta_1 x} - \dfrac{1}{c_1^2\zeta_1} \displaystyle\int_{-\infty}^{x} f(x') \sin[\zeta_1(x - x')] dx' & \textit{si } x < 0, \\ (\mathcal{C}_n(\zeta)f)\, e^{i\zeta_2 x} + \dfrac{1}{c_2^2\zeta_2} \displaystyle\int_{x}^{+\infty} f(x') \sin[\zeta_2(x - x')] dx' & \textit{si } x > 0, \end{cases}$$

avec

$$\mathcal{B}_n(\zeta)f = \frac{1}{(c_1^2\zeta_1 + c_2^2\zeta_2)} \left\{ \begin{array}{l} \displaystyle\int_{-\infty}^{0} [-\frac{c_2^2\zeta_2}{c_1^2\zeta_1} \sin(\zeta_1 x') + i\cos(\zeta_1 x')] f(x') dx' \\ + \\ \displaystyle\int_{0}^{+\infty} [-\sin(\zeta_2 x') + i\cos(\zeta_2 x')] f(x') dx', \end{array} \right.$$

$$\mathcal{C}_n(\zeta)f = \frac{1}{(c_1^2\zeta_1 + c_2^2\zeta_2)} \left\{ \begin{array}{l} \displaystyle\int_{-\infty}^{0} [\sin(\zeta_1 x') + i\cos(\zeta_1 x')] f(x')dx' \\ + \\ \displaystyle\int_{0}^{+\infty} [\frac{c_1^2\zeta_1}{c_2^2\zeta_2}\sin(\zeta_2 x') + i\cos(\zeta_2 x')] f(x')dx'. \end{array} \right.$$

Etape E3. La formule de Stone. Elle met en relation la famille des projecteurs spectraux $\Pi_n(a)$, a réel, avec les résolvantes $R_{B_n}(\zeta)$, ζ nombre complexe n'appartenant pas au spectre $\sigma(B_n)$. Dans notre cas où $\sigma(B_n) = \sigma_c(B_n)$, elle s'écrit

$$\forall a, b \in \mathbb{R},\ \forall f, g \in L^2(\mathbb{R}),\quad (\Pi_n(b)f/g)_{L^2(\mathbb{R})} - (\Pi_n(a)f/g)_{L^2(\mathbb{R})} \tag{4.10}$$

$$= \frac{1}{2i\pi} \lim_{\varepsilon \to 0_+} \int_a^b \left[(R_{B_n}(\lambda + i\varepsilon)f/g)_{L^2(\mathbb{R})} - (R_{B_n}(\lambda - i\varepsilon)f/g)_{L^2(\mathbb{R})} \right] d\lambda.$$

A partir de l'expression de la résolvante donnée au théorème 4.2 , on peut calculer la limite de l'intégrale figurant au second membre de (4.10). Ce calcul fait apparaître les fonctions propres généralisées $\Phi^i(\lambda, n, x)$, $i = 0, 1$ ou 2, définies en (4.3), (4.5) et (4.6).

THÉORÈME 4.3. *On se donne les fonctions f et g dans $C_0^\infty(\mathbb{R})$, et l' intervalle $[a, b] \subset \mathbb{R}$ avec $a < c_2^2 q_n^2 < b$. Alors*

$$\frac{1}{2i\pi} \lim_{\varepsilon \to 0_+} \int_a^b \left[(R_{B_n}(\lambda + i\varepsilon)f/g)_{L^2(\mathbb{R})} - (R_{B_n}(\lambda - i\varepsilon)f/g)_{L^2(\mathbb{R})} \right] d\lambda$$

$$= \frac{1}{2i\pi} \int_{c_2^2 q_n^2}^b \left[(R_{B_n}(\lambda + i0)f/g)_{L^2(\mathbb{R})} - (R_{B_n}(\lambda - i0)f/g)_{L^2(\mathbb{R})} \right] d\lambda,$$

avec

$$\frac{1}{2i\pi} \left[(R_{B_n}(\lambda + i0)f/g)_{L^2(\mathbb{R})} - (R_{B_n}(\lambda - i0)f/g)_{L^2(\mathbb{R})} \right] \tag{4.11}$$

$$= \left\{ \begin{array}{ll} \dfrac{1}{\pi} \dfrac{c_2^2\xi_2}{(c_1^2\xi_1')^2 + (c_2^2\xi_2)^2} (f/\Phi^0)(\overline{g/\Phi^0}) & si\ \lambda \in I_n = (c_2^2 q_n^2, c_1^2 q_n^2), \\ \dfrac{1}{\pi} \left[\begin{array}{l} \dfrac{c_2^2\xi_2}{c_1^2\xi_1(c_1^2\xi_1 + c_2^2\xi_2)} (f/\Phi^1)(\overline{g/\Phi^1}) \\ + \\ \dfrac{1}{c_1^2\xi_1 + c_2^2\xi_2} (f/\Phi^2)(\overline{g/\Phi^2}) \end{array} \right] & si\ \lambda \in J_n = (c_1^2 q_n^2, +\infty), \end{array} \right.$$

formule dans laquelle la notation (f/Φ) désigne l'intégrale de $f\overline{\Phi}$ sur $\mathbb{R}$.

Etape E4. Normalisation des fonctions propres généralisées et représentation spectrale de B_n. La formule de Stone et les formules du théorème 4.3 conduisent à choisir les coefficients $a^i(\lambda, n)$ introduits au théorème 4.1 tels que

$$a^0(\lambda, n) = \left(\frac{c_2^2\xi_2}{\pi[(c_1^2\xi_1')^2 + (c_2^2\xi_2)^2]} \right)^{1/2}, \tag{4.12a}$$

$$a^1(\lambda,n)=\left(\frac{c_2^2\xi_2}{\pi c_1^2\xi_1(c_1^2\xi_1+c_2^2\xi_2)}\right)^{1/2}, \tag{4.12b}$$

$$a^2(\lambda,n)=\left(\frac{1}{\pi(c_1^2\xi_1+c_2^2\xi_2)}\right)^{1/2}, \tag{4.12c}$$

et on obtient ainsi les fonctions propres

$$\varphi^0=\varphi^0(\lambda,n,x),\quad \varphi^1=\varphi^1(\lambda,n,x)\quad\text{et}\quad \varphi^2=\varphi^2(\lambda,n,x)$$

des formules (2.19a), (2.19b), et (2.19c) du théorème 2.3 du §2.

On peut alors énoncer les propositions qui suivent.

PROPOSITION 4.2. *Pour toute fonction f de $L^2(\mathbb{R})$, la limite suivante*

$$\widetilde{f}^0(\lambda,n)=L^2(I_n\,;d\lambda)-\lim_{N\to+\infty}\int_{-N}^{N}f(x)\overline{\varphi^0(\lambda,n,x)}dx \tag{4.13}$$

existe et définit un coefficient de Fourier généralisé pour f, c'est-à-dire que

$$\widetilde{\mathcal{F}}_n^0\begin{cases} L^2(\mathbb{R}) & \longrightarrow\ \widetilde{\mathcal{H}}_n^0=L^2(I_n)\\ f & \longmapsto\ \widetilde{f}^0(.,n)\end{cases} \tag{4.14}$$

est un opérateur partiellement isométrique de sous-espace initial $\Pi_n(\overline{I_n})L^2(\mathbb{R})=\mathcal{H}_{1,n}^0$ et de sous-espace final $\widetilde{\mathcal{H}}_n^0$. L'application réciproque est donnée par

$$(\widetilde{\mathcal{F}}_n^0)^*(\widetilde{f}^0)(x)=L^2(\mathbb{R})-\lim_{\delta\to0_+}\int_{c_2^2q_n^2+\delta}^{c_1^2q_n^2-\delta}\widetilde{f}^0(\lambda,n)\varphi^0(\lambda,n,x)d\lambda. \tag{4.15}$$

L'application $\widetilde{\mathcal{F}}_n^0$ est une représentation spectrale de la restriction de B_n à $\mathcal{H}_{1,n}^0$, sur l'espace $\widetilde{\mathcal{H}}_n^0$, qui transforme cette restriction en opérateur de multiplication par λ dans $\widetilde{\mathcal{H}}_n^0$, c'est-à-dire

$$\widetilde{\mathcal{F}}_n^0(B_nf)(\lambda)=\lambda\widetilde{f}^0(\lambda,n). \tag{4.16}$$

Cet opérateur, $B_n|_{\mathcal{H}_{1,n}^0}$ est absolument continu de spectre $\overline{I_n}=[c_2^2q_n^2,c_1^2q_n^2]$.

PROPOSITION 4.3. *Pour toute fonction f de $L^2(\mathbb{R})$, pour $i=1$ ou 2, la limite suivante*

$$\widetilde{f}^i(\lambda,n)=L^2(J_n\,;d\lambda)-\lim_{N\to+\infty}\int_{-N}^{N}f(x)\overline{\varphi^i(\lambda,n,x)}dx \tag{4.17}$$

existe et permet de définir un opérateur $\widetilde{\mathcal{F}}_n^1$ partiellement isométrique de sous-espace initial $\Pi_n(\overline{J_n})L^2(\mathbb{R})=\mathcal{H}_{1,n}$ et de sous-espace final $L^2(J_n)\oplus L^2(J_n)=\widetilde{\mathcal{H}}_n^1\oplus\widetilde{\mathcal{H}}_n^2$, tel que

$$\widetilde{\mathcal{F}}_n^1\begin{cases}\mathcal{H}_{1,n}\to\widetilde{\mathcal{H}}_n^1\oplus\widetilde{\mathcal{H}}_n^2\\ f\longmapsto(\widetilde{f}^1(\lambda,n),\widetilde{f}^2(\lambda,n)),\end{cases} \tag{4.18}$$

$$(\widetilde{\mathcal{F}}_n^1)^*(\widetilde{f}^1,\widetilde{f}^2)(x)=L^2(\mathbb{R})-\lim_{\substack{\delta\to0_+\\N\to+\infty}}\sum_{i=1}^{2}\int_{c_1^2q_n^2+\delta}^{N}\widetilde{f}^i(\lambda,n)\varphi^i(\lambda,n,x)d\lambda, \tag{4.19}$$

$$\widetilde{\mathcal{F}}_n^1(B_n f)(\lambda) = (\lambda \widetilde{f}^1(\lambda, n), \lambda \widetilde{f}^2(\lambda, n)). \tag{4.20}$$

$\widetilde{\mathcal{F}}_n^1$ définit une représentation spectrale de la restriction de B_n à $\mathcal{H}_{1,n}$, qui est un opérateur absolument continu de spectre $\overline{J_n} = [c_1^2 q_n^2, +\infty)$.

PROPOSITION 4.4 (Projecteurs spectraux de l'opérateur B_n). *Soient b un nombre réel et f une fonction de $L^2(\mathbb{R})$. Alors ou $b \leq c_2^2 q_n^2$ et $\Pi_n(b)(f) = 0$, ou*

$$b \in \overline{I_n} = [c_2^2 q_n^2, c_1^2 q_n^2] \quad et \quad \Pi_n(b)(f) = \int_{c_2^2 q_n^2}^{b} \widetilde{f}^0(\lambda, n)\varphi^0(\lambda, n, .)d\lambda,$$

ou

$$b \in \overline{J_n} = [c_1^2 q_n^2, +\infty) \text{ et } \Pi_n(b)(f) = \int_{I_n} \widetilde{f}^0(\lambda, n)\varphi^0(\lambda, n, .)d\lambda + \sum_{i=1}^{2} \int_{J_n} \widetilde{f}^i(\lambda, n)\varphi^i(\lambda, n, .)d\lambda.$$

On peut maintenant expliciter une représentation spectrale de l'opérateur B_n qui est utilisée au théorème 2.4.

THÉORÈME 4.4 (Représentation spectrale de l'opérateur B_n). *Soit*

$$\widetilde{\mathcal{H}}_n = \widetilde{\mathcal{H}}_n^0 \oplus \widetilde{\mathcal{H}}_n^1 \oplus \widetilde{\mathcal{H}}_n^2 = L^2(I_n) \oplus L^2(J_n) \oplus L^2(J_n) \tag{4.21}$$

la somme directe hilbertienne des espaces introduits aux propositions 4.2 et 4.3. La transformation unitaire

$$\widetilde{\mathcal{F}}_n \begin{cases} L^2(\mathbb{R}) \to \widetilde{\mathcal{H}}_n \\ f \longmapsto (\widetilde{f}^0(\lambda, n), \widetilde{f}^1(\lambda, n), \widetilde{f}^2(\lambda, n)) \end{cases} \tag{4.22}$$

où les $\widetilde{f}^i(\lambda, n), i = 0, 1, 2$, sont définis en (4.13) et (4.17), réduit l'opérateur B_n. Pour toute fonction f de $L^2(\mathbb{R})$, l'égalité de Bessel–Parseval s'écrit

$$\|f\|^2_{L^2(\mathbb{R})} = \|\widetilde{f}^0(., n)\|^2_{L^2(I_n)} + \|\widetilde{f}^1(., n)\|^2_{L^2(J_n)} + \|\widetilde{f}^2(., n)\|^2_{L^2(J_n)} \tag{4.23}$$

et la transformation $\widetilde{\mathcal{F}}_n^{-1}$ permet d'écrire le développement

$$f(x) = \int_{I_n} \widetilde{f}^0(\lambda, n)\varphi^0(\lambda, n, x)d\lambda + \sum_{i=1}^{2} \int_{J_n} \widetilde{f}^i(\lambda, n)\varphi^i(\lambda, n, x)d\lambda \tag{4.24}$$

dans lequel les intégrales en λ convergent au sens de la norme de $L^2(\mathbb{R}; dx)$.

4.2. L'opérateur B_ξ sur $L^2((0, H))$. On considère un profil $c(z)$ qui satisfait l'hypothèse

$$\textbf{(H)} \qquad c \in L^\infty((0, H)) \quad \text{et} \quad \operatorname{Min} c(z) \geq c_m > 0.$$

Pour ξ réel, l'opérateur réduit B_ξ a été introduit en (3.6). Son domaine $D(B)$ est indépendant de ξ. Nous les rappelons :

$$B_\xi = -\frac{d}{dz}\left(c^2(z)\frac{d}{dz}\right) + c^2(z)\xi^2, \tag{4.25}$$

$$D(B) = \{u \in H^1((0, H)) \,/\, c^2 u' \in H^1((0, H)) \text{ et } (c^2 u')(0) = u(H) = 0\}.$$

C'est un opérateur auto-adjoint dans $L^2((0,H))$, minoré et à inverse compact.

Soit $(\lambda(\xi,n))_{n\geq 1}$ la suite des valeurs propres de B_ξ, rangées par ordre croissant et répétées éventuellement avec leur ordre de multiplicité. Soit $(V(\xi,n,.))_{n\geq 1}$ une base orthonormée de $L^2((0,H))$, formée de fonctions propres associées. On a donc

$$V(\xi,n,.)\in D(B) \text{ et } B_\xi(V(\xi,n,.)) = \lambda(\xi,n)V(\xi,n,.). \tag{4.26}$$

Donnons quelques propriétés de ces fonctions propres et valeurs propres.

THÉORÈME 4.5. *Avec un profil $c(z)$ vérifiant* **(H)** *et avec les notations ci-dessus, on a les propriétés suivantes.*

(a) *Chaque valeur propre $\lambda(\xi,n)$ est simple.*

(b) *Chaque fonction $\lambda(.,n)$, définie sur $\mathbb{R}$, est analytique, paire, strictement croissante sur l'intervalle $[0,+\infty)$ et à valeurs dans l'intervalle $\overline{I}_n = [\lambda(0,n),+\infty)$. Sa dérivée première est donnée par*

$$\partial_\xi\lambda(\xi,n) = 2\xi\int_0^H c^2(z)\,|V(\xi,n,z)|^2dz. \tag{4.27}$$

On a en particulier $\partial_\xi^2\lambda(0,n)\neq 0$, et

$$\lambda(\xi,n) = \lambda(0,n)+\xi^2a_n(\xi) \text{ avec } c_m^2\leq a_n(\xi)\leq \|c\|^2_{L^\infty((0,H))}\,. \tag{4.28}$$

(c) *On peut choisir la fonction propre $V(\xi,n,.)$ telle que la fonction $\xi\mapsto V(\xi,n,.)$ définie sur $\mathbb{R}$ à valeurs dans $L^2((0,H))$ soit réelle analytique et paire.*

Preuve. Soient $V_1(\xi,n,.)$ et $V_2(\xi,n,.)$ deux fonctions propres associées à $\lambda(\xi,n)$. La fonction $c^2\partial_zV_1V_2 - V_1c^2\partial_zV_2$ est dans $H^1((0,H))$, à dérivée nulle, et nulle en $z=0$ ou $z=H$. Elle est donc nulle sur $(0,H)$. On en déduit la colinéarité locale de V_1 et V_2. Un argument de connexité donne la colinéarité globale sur $(0,H)$.

On a

$$\lambda(\xi,n) = (\,B_\xi V(\xi,n,.)\,/\,V(\xi,n,.)\,)_{L^2((0,H))}$$

$$= \int_0^H c^2(z)\,|\partial_zV(\xi,n,z)|^2dz + \xi^2\int_0^H c^2(z)\,|V(\xi,n,z)|^2dz. \tag{4.29}$$

On en déduit

$$\lambda(\xi,n) > \xi^2\,\mathrm{Min}\,c^2(z) \text{ et } \lim_{\xi\to+\infty}\lambda(\xi,n) = +\infty. \tag{4.30}$$

La famille $(B_\xi)_{\xi\in\mathbb{R}}$ d'opérateurs à résolvante compacte, est une famille analytique auto-adjointe de type (A) au sens de Kato (cf. [Ka, chapitre VII, §2]). Le théorème 3.9 de ce même chapitre de [Ka] donne l'analyticité et la parité des fonctions $\xi\to\lambda(\xi,n)$ et $\xi\to V(\xi,n,.)$.

La formule du Min-Max (cf. [DS]) pour la $n$ième valeur propre $\lambda(\xi,n)$ de l'opérateur B_ξ, ξ réel, permet d'établir la croissance de la fonction $\lambda(.,n)$ sur $[0,+\infty)$. L'analyticité et (4.30) entraînent la stricte croissance sur ce même intervalle.

La fonction $\partial_\xi V(\xi,n,.)$ est dans le domaine $D(B)$ de l'opérateur B_0. En effet

$$W_{\xi'}(\xi,n,.) = \frac{V(\xi',n,.)-V(\xi,n,.)}{\xi'-\xi}\ \in\ D(B),$$

$$B_0 W_{\xi'}(\xi,n,.) = \frac{B_0 V(\xi',n,.) - B_0 V(\xi,n,.)}{\xi' - \xi}$$

$$= \frac{\lambda(\xi',n)V(\xi',n,.) - \lambda(\xi,n)V(\xi,n,.) - c^2(\,\xi'^2 V(\xi',n,.) - \xi^2 V(\xi,n,.)\,)}{\xi' - \xi}.$$

Quand ξ' tend vers ξ, on a la convergence dans $L^2((0,H))$ de la famille $(\,W_{\xi'}(\xi,n,.)\,)_{\xi'}$ vers $\partial_\xi V(\xi,n,.)$ et de la famille $(\,B_0 W_{\xi'}(\xi,n,.)\,)_{\xi'}$ vers

$$Z(\xi,n,.) = \partial_\xi\lambda(\xi,n)V(\xi,n,.) + \lambda(\xi,n)\partial_\xi V(\xi,n,.) - c^2(\,2\xi V(\xi,n,.) + \xi^2\partial_\xi V(\xi,n,.)\,).$$

Le point $(\partial_\xi V(\xi,n,.), Z(\xi,n,.))$ est donc adhérent au graphe de l'opérateur B_0 et

$$B_\xi\,\partial_\xi V(\xi,n,.) = \partial_\xi\lambda(\xi,n)V(\xi,n,.) + \lambda(\xi,n)\partial_\xi V(\xi,n,.) - 2\xi c^2 V(\xi,n,.).$$

Par ailleurs, le produit scalaire avec $V(\xi,n,.)$ s'écrit

$$(\,B_\xi\partial_\xi V(\xi,n,.)\,/\,V(\xi,n,.)\,)_{L^2((0,H))} = (\,\partial_\xi V(\xi,n,.)\,/\,B_\xi V(\xi,n,.)\,)_{L^2((0,H))}$$

$$= \lambda(\xi,n)(\,\partial_\xi V(\xi,n,.)\,/\,V(\xi,n,.)\,)_{L^2((0,H))}.$$

On en déduit la dérivée (4.27) et la forme (4.28) pour $\lambda(.,n)$. □

La figure 4.2 donne l'allure des courbes de dispersion $\xi \to \lambda(\xi,n)$, qui peuvent éventuellement présenter des points d'inflexion, puisque la dérivée seconde de $\lambda(.,n)$ s'écrit

$$\partial_\xi^2\lambda(\xi,n) = 2\int_0^H c^2(z)\,|V(\xi,n,z)|^2 dz \tag{4.31}$$

$$+\,4\,\xi\int_0^H c^2(z)\,\mathrm{Re}\,(\,\partial_\xi V(\xi,n,z)\overline{V(\xi,n,z)}\,)dz.$$

Remarque 4.2. Toutes ces propriétés restent vraies pour l'opérateur B_ξ, avec des conditions limites de Dirichlet, respectivement Neumann, en $z = 0$ et $z = H$. Le domaine $D(B) = \{u \in H^1((0,H))\,/\,c^2u' \in H^1((0,H)),\ u(0) = u(H) = 0\}$, respectivement $D(B) = \{u \in H^1((0,H))\,/\,c^2u' \in H^1((0,H)),\ (c^2u')(0) = (c^2u')(H) = 0\}$, reste indépendant de ξ.

Considérons maintenant le cas particulier d'une bande présentant deux strates horizontales. Le profil c est alors (1.6) et représenté à la figure 4.2: h est un réel de $[0,H]$, c_1 et c_2 sont deux réels strictement positifs, $c(z)$ est égal à c_1 si $z \in (0,h)$ et à c_2 si $z \in (h,H)$.

Les valeurs propres sont solutions de la relation de dispersion, qui traduit les conditions de raccord en $z = h$ des solutions V_1 et V_2 du système

$$\begin{cases} -c_1^2 V_1'' + c_1^2\xi^2 V_1 = \lambda V_1 \text{ sur } (0,h), \\ -c_2^2 V_2'' + c_2^2\xi^2 V_2 = \lambda V_2 \text{ sur } (h,H), \\ V_1'(0) = V_2(H) = 0, \\ V_1(h) = V_2(h) \text{ et } c_1^2 V_1'(h) = c_2^2 V_2'(h). \end{cases}$$

Cette relation s'écrit

$$\tan\left[\frac{h}{c_1}(\lambda - c_1^2\xi^2)^{1/2}\right]\tan\left[\frac{(H-h)}{c_2}(\lambda - c_2^2\xi^2)^{1/2}\right] = \frac{c_2(\lambda - c_2^2\xi^2)^{1/2}}{c_1(\lambda - c_1^2\xi^2)^{1/2}}. \tag{4.32}$$

Les solutions réelles de (4.32), rangées par ordre croissant, déterminent la suite $(\lambda(\xi,n))_{n\in\mathbb{N}^*}$. L'équation donnant la suite des seuils $\lambda(0,n)$ s'écrit

$$\tan\left(\frac{h}{c_1}\lambda^{1/2}\right)\ \tan\left[\frac{(H-h)}{c_2}\lambda^{1/2}\right]=\frac{c_2}{c_1}. \tag{4.33}$$

La direction propre associée à $\lambda=\lambda(\xi,n)$ est donnée par

$$V(\xi,n,z)=\begin{cases} A(\xi,n)\cos[\xi_1(\lambda)z] \text{ si } z\in(0,h),\\ A(\xi,n)\dfrac{\cos[\xi_1(\lambda)h]}{\sin[\xi_2(\lambda)(H-h)]}\sin[\xi_2(\lambda)(H-z)] \text{ si } z\in(h,H), \end{cases} \tag{4.34}$$

avec

$$\xi_1(\lambda)=\frac{1}{c_1}(\lambda-c_1^2\xi^2)^{1/2} \text{ et } \xi_2(\lambda)=\frac{1}{c_2}(\lambda-c_2^2\xi^2)^{1/2},$$

$$\operatorname{Im}\xi_j(\lambda)\geq 0 \text{ si } \lambda\leq c_j^2\xi^2, \qquad j=1 \text{ ou } 2.$$

La condition de normalisation détermine les coefficients $A=A(\xi,n)$. En posant $\xi_j=\xi_j(\lambda(\xi,n))$, on a

$$\frac{A^2}{2}\left[h+\frac{\sin(2\xi_1h)}{2\xi_1}+(H-h)\frac{\cos^2(\xi_1h)}{\sin^2[\xi_2(H-h)]}-\frac{c_1^2\xi_1\sin(2\xi_1h)}{2c_2^2\xi_2^2}\right]=1. \tag{4.35}$$

Examinons enfin le cas particulier $c_1=c_2=c$. Une résolution directe de (4.26) donne

$$\begin{cases} n\geq 1,\ q_n=\dfrac{1}{H}[\dfrac{\pi}{2}+(n-1)\pi] \text{ et } \lambda(\xi,n)=c^2q_n^2+c^2\xi^2,\\ V(\xi,n,z)=\sqrt{\dfrac{2}{H}}\sin[q_n(H-z)] \quad (=(-1)^n\sqrt{\dfrac{2}{H}}\cos(q_nz)\ ), \end{cases} \tag{4.36}$$

et on retrouve les fonctions (2.9). Les valeurs propres et fonctions propres peuvent aussi être obtenues comme cas limite de (4.32) et (4.34), puisque (4.32) s'écrit $\tan(h\xi)\ \tan[(H-h)\xi]=1$, avec $\xi=\xi_1(\lambda)=\xi_2(\lambda)=\dfrac{1}{c}(\lambda-c^2\xi^2)^{1/2}$. On obtient

$$\xi=\frac{1}{H}\left[\frac{\pi}{2}+(n-1)\pi\right]=q_n, \qquad n\geq 1.$$

Les courbes de dispersion $\xi\to\lambda(\xi,n)=c^2q_n^2+c^2\xi^2$ sont des paraboles passant pour $\xi=0$ par les seuils $\lambda(0,n)=c^2q_n^2$.

4.3. Figures. Le profil (1.5) avec $c_1>c_2$, donne le spectre (2.23), soit

$$\sigma(A)=\left[\frac{c_2^2\pi^2}{4H^2},\ +\infty\right)=\bigcup_{n\geq 1}\sigma(B_n)=\bigcup_{n\geq 1}[c_2^2q_n^2,+\infty), \text{ avec } q_n=\frac{1}{H}[\frac{\pi}{2}+(n-1)\pi].$$

Nous le visualisons à l'aide des spectres (4.7) des opérateurs réduits B_n (Fig. 4.1).

Le profil (1.6) donne le spectre (3.18), soit

$$\sigma(A)=[\lambda(0,1),+\infty)=\bigcup_{\xi\in\mathbb{R}}\sigma(B_\xi)=\{\lambda(\xi,n)\ /\ \xi\in\mathbb{R} \text{ et } n\in\mathbb{N}^*\}.$$

Selon (4.28), chaque courbe de dispersion $\xi\to\lambda(\xi,n)$ est située dans une bande limitée par les deux paraboles d'équations

$$\xi\to\lambda(0,n)+c_m^2\xi^2 \text{ et } \xi\to\lambda(0,n)+c_M^2\xi^2,$$

où $c_m=\operatorname{Min}c(z)$ et $c_M=\operatorname{Max}c(z)$; voir Fig. 4.2.

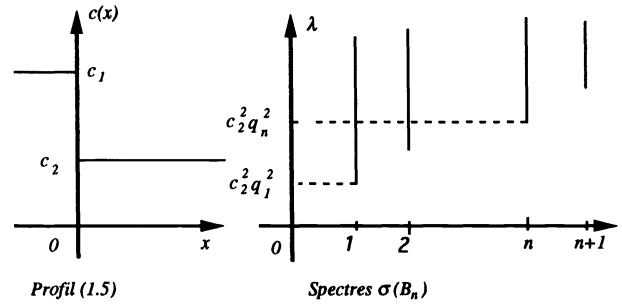


FIG. 4.1.

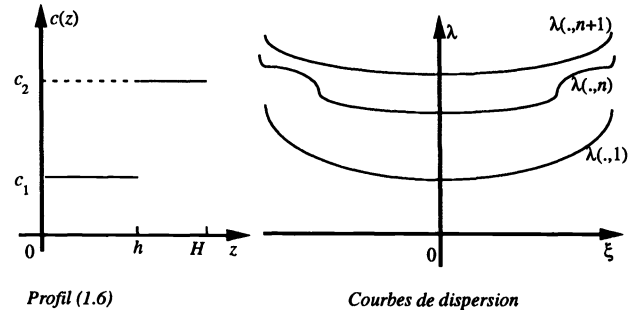


FIG. 4.2.

Remerciements. Les auteurs remercient Jean-Luc Boelle, ingénieur à Elf Aquitaine Production, qui leur a soumis ce problème, ainsi que le professeur Hiroshi Isozaki de l'Université d'Oşaka pour les discussions fructueuses qu'ils ont eues sur ce travail et les modifications qu'il suggéra, lors de sa visite à l'Université de Provence au mois de septembre 1992. Les auteurs remercient aussi les rapporteurs de cet article, pour leurs remarques judicieuses, qui ont permis d'établir des correspondances avec d'autres travaux et de compléter certains points.

REFERENCES

[A] S. AGMON, *Spectral properties of Schrödinger operators and scattering theory*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), II (1975), pp. 151–218.

[B] J. L. BOELLE, *Modélisation numérique de la propagation sismique à partir d'un puits : Méthode modale pour l'étude de la diffraction des ondes de tube*, Rapport SNEA (P) pour le Projet "Sismique entre puits: nouvelles recherches," Ref. CST 355: F 35346, Mars 1992.

[BD] M. BEN ARTZI ET A. DEVINATZ, *The limiting absorption principle for partial differential operators*, Mem. Amer. Math. Soc., 66 (1987), n^{0}364, pp. 1–70.

[BDG] M. BEN ARTZI, Y. DERMENJIAN, ET J .C. GUILLOT, *Acoustic waves in perturbed stratified fluids : a spectral theory*, Comm. Partial Differential Equations, 14 (1989), pp. 479–517.

[CD] E. CROC ET Y. DERMENJIAN, *Principe d'absorption limite et fonctions de Green pour un problème de sismique dans un milieu stratifié dans une direction*, Rapport interne n° 2, février 1993, J. E. 180 EDP-AN, Université de Provence (Marseille), France.

[DG86] Y. DERMENJIAN ET J .C. GUILLOT, *Théorie spectrale de la propagation des ondes acoustiques dans un milieu stratifié perturbé*, J. Differential Equations, 62 (1986), pp. 357–409.

[DG88] ———, *Scattering of elastic waves in a perturbed isotropic half space with a free boundary. The limiting absorption principle*, Math. Methods Appl. Sci. 10 (1988), pp. 87–124.

[DL] R. DAUTRAY ET J .L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology, Vol. 3, Spectral Theory and Applications*, Springer-Verlag, Berlin, 1990.

[DP] S. DEBIÈVRE ET D .W. PRAVICA, *Spectral analysis for optical fibres and stratified fluids.* I. *The limiting absorption principle*, J. Funct. Anal. 98 (1991), pp. 404–436.

[DS] N. DUNFORD ET J .T. SCHWARTZ, *Linear operators. Part* II. *Spectral theory. Selfadjoint operators in Hilbert space*, Wiley–Interscience, New York, London, 1963.

[E69] D. EIDUS, *The principle of limiting amplitude*, Russ. Math. Surv., 24 (1969), pp. 97–167.

[E86] ———, *The limiting absorption and amplitude principles for the diffraction problem with two unbounded media*, Comm. Math. Phys., 107 (1986), pp. 29–38.

[G] J .C. GUILLOT, *Complétude des modes T.E. et T.M. pour un guide d'ondes optiques planaire*, Rapport de recherche INRIA n° 385, 1985.

[GW] J .C. GUILLOT ET C .H. WILCOX, *Spectral analysis of the Epstein operator*, Proc. Roy. Soc. Edinburgh Sect. A, 80 (1978), pp. 85–98.

[Ha] G. HACHEM, *Spectral theory for Dirac operators with a Stark potential*, J. Math. Pures Appl., 71 (1992), pp. 293–329.

[Hö] L. HÖRMANDER, *The analysis of linear partial differential operators. II. Differential operators with constant coefficients*, Grundelehren der mathematischen Wissenschaften, 257, Springer-Verlag, Berlin, New York, 1983.

[Ka] T. KATO, *Perturbation theory for linear operators*, Grundelehren der mathematischen Wissenschaften, 132, Springer-Verlag, Berlin, New York, 1976.

[KT] K. KIKUCHI ET H. TAMURA, *Limiting amplitude principle for acoustic propagators in perturbed stratified fluids*, J. Differential Equations, 93 (1991), pp. 260–282.

[L] W .C. LYFORD, *Spectral analysis of the Laplacian in domains with cylinders*, Math. Ann., 218 (1975), pp. 229–251.

[MW87] K. MORGENRÖTHER ET P. WERNER, *Resonances and standing waves*, Math. Methods Appl. Sci. 9 (1987), n° 1, pp. 105–126.

[MW88] ———, *On the principles of limiting absorption and limit amplitude for a class of locally perturbed waveguides.* I. *Time-independent theory*, Math. Methods Appl. Sci., 10 (1988), pp. 125–144.

[RS] M. REED ET B. SIMON, *Methods of Modern Mathematical Physics.* I. *Functional Analysis*, Academic Press, New York, London, 1972.

[SZ] J. SJÖSTRAND ET M. ZWORSKI, *Complex scaling and the distribution of scattering poles*, J. Amer. Math. Soc., 4 (1991), pp. 729–769.

[T] H. TAMURA, *The principle of limiting absorption for propagative systems in crystal optics with perturbations of long range class*, Nagoya Math. J., 84 (1981), pp. 169–193.

[W86] P. WERNER, *Low frequency asymptotics for the reduced wave equation in two-dimensional exterior domains*, Math. Methods Appl. Sci., 8 (1986), pp. 134–156.

[W87] ———, *Resonance phenomena in cylindrical waveguides*, J. Math. Anal. Appl., 121 (1987), pp. 173–214.

[We] R. WEDER, *Spectral and Scattering Theory for Wave Propagation in Perturbed Stratified Media*, Applied Mathematical Sciences, 87, Springer-Verlag, New York, Berlin, 1991.

[Wi75] C .H. WILCOX, *Scattering theory for the D'Alembert equation in exterior domains*, Lecture Notes in Math., Vol. 442, Springer-Verlag, Berlin, New York, 1975.

[Wi84] ———, *Sound propagation in stratified fluids*, Applied Mathematical Sciences, 50, Springer-Verlag, New York, Berlin, 1984.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 925–943, July 1995

007

PYRAMIDAL ALGORITHMS FOR LITTLEWOOD–PALEY DECOMPOSITIONS*

M. A. MUSCHIETTI† AND B. TORRÉSANI‡

Abstract. It is well known that a pyramidal algorithm is associated with any usual multiresolution analysis of $L^2(IR)$ for the computation of the corresponding wavelet coefficients. It is shown that an approximate pyramidal algorithm may be associated with more general Littlewood–Paley decompositions. Accuracy estimates are provided for such approximate algorithms. Finally, some explicit examples are studied.

Key words. wavelets, subband coding

AMS subject classifications. 47A58, 42C15, 66AO5, 66A35

1. Introduction.

Wavelet analysis has emerged in the past ten years as a completely generic methodology for solving problems in many different areas such as mathematical analysis and operator theory, numerical analysis, signal and image processing, computer vision, computer music, turbulence, and astrophysics. Among the advantages of wavelet decompositions, their relative simplicity and the existence of associated fast algorithms are two of the most important [1], [6].

Essentially there exist two different approaches to wavelets, namely, the discrete and the continuous approaches. Roughly speaking, discrete wavelet decompositions are most often adapted to problems in which it is important to reduce the volume of data, for instance, in signal or image compression or numerical analysis. On the other hand, for physical signal analysis problems, one is interested in keeping redundancy on the wavelet transform to get a finer analysis.

The main drawback of continuous wavelet decompositions is that there is a priori no associated fast algorithm for the computation of the corresponding wavelet transform. Some attempts have been made to cure such a drawback, mainly by matching a multiresolution framework to the continuous setting (see, for instance, [4], [7]). They are, in general, associated with limited classes of wavelets.

Here we describe a method for associating fast algorithms to continuous wavelet decomposition, based on the same philosophy. In particular, it is shown that, starting from a usual mother wavelet, the scale discretization yields a new wavelet (called the integrated wavelet) which is associated with a pair of low- and high-pass filters. These filters are, in general, not discrete, but may in some situations be well approximated by discrete filters, the localization of which can be directly related to the regularity of the scaling function.

The paper is organized as follows: Section 2 is devoted to a description of the version of continuous wavelet decompositions that we will use in what follows. In §3 we present the algorithmic aspects we are interested in and, in particular, our main result, Theorem 3.2. We present some examples in §4, and §5 is devoted to the conclusion.

* Received by the editors July 16, 1993; accepted for publication December 15, 1993. This research was supported in part by a Centre National de la Recherche Scientifique (Paris)/CONICET cooperation agreement and the Physics department of Université de Provence, Marseille.

† Departamento de Matemática, Universidad de La Plata, Argentina.

‡ Centre de Physique Theorique, Centre National de la Recherche Scientifique Luminy, Case 907, F-13288 Marseille, Cedex 09, France.

Throughout this paper we shall use the following notation. We shall denote by $||\cdot||_p$ the $L^p(\mathbb{R})$, $L^p([-\pi,\pi])$, and $\ell^p(\mathbb{Z})$ norms. The $L^p([-\pi,\pi])$ norm is normalized as

$$||f||_p = \left(\int_{[-\pi,\pi]} |f(x)|^p dx \right)^{\frac{1}{p}}.$$

Our conventions for the Hermitian product and Fourier transform in $L^2(\mathbb{R})$ are the following ones

$$\langle f, g \rangle = \int f(x) g(x)^* \, dx,$$

where the star denotes complex conjugation, and

$$\hat{f}(\xi) = \int_{\mathbb{R}} f(x) e^{-i\xi x} dx.$$

2. Continuous wavelet decompositions. Let us start from standard notions of continuous wavelet analysis. We will focus on the analysis of $L^2(\mathbb{R})$, and sometimes we will describe in a few words the corresponding results in the $H^2(\mathbb{R})$ context (here we will denote the complex Hardy space by $H^2(\mathbb{R}) = \{f \in L^2(\mathbb{R}), \hat{f}(\xi) = 0 \ \forall \xi \leq 0\}$). Here we shall be interested in two "decomposition-reconstruction" schemes corresponding to different "reconstruction wavelets."

2.1. The bilinear scheme. Generically, an *infinitesimal wavelet* (or mother wavelet) is a function $\psi \in L^1(\mathbb{R})$ such that the following admissibility condition holds:

$$c_\psi = \int_0^\infty \left|\hat{\psi}(u)\right|^2 \frac{du}{u} = \int_0^\infty \left|\hat{\psi}(-u)\right|^2 \frac{du}{u} = 1 \tag{1}$$

(in such a case, ψ is generally taken to be a real-valued function). If $\hat{\psi}$ is, say, differentiable, equation (1) basically means that $\hat{\psi}(0) = 0$, which can be otherwise stated as follows:

$$\int_{-\infty}^{\infty} \psi(x) \, dx = 0. \tag{2}$$

Such a mother wavelet provides the following analysis of $L^2(\mathbb{R})$: for any $(b,a) \in \mathbb{R} \times \mathbb{R}_+^*$, one introduces the wavelet

$$\psi_{(b,a)}(x) = \frac{1}{a} \, \psi\left(\frac{x-b}{a}\right), \tag{3}$$

and one has the following representation theorem, the proof of which is well known and can be found in [3] or [2].

THEOREM 2.1 (Calderón). *Let ψ be a mother wavelet. Then any $f \in L^2(\mathbb{R})$ decomposes as follows:*

$$\begin{aligned} f &= \int_{\mathbb{R}\times\mathbb{R}_+^*} T_f(b,a) \, \psi_{(b,a)} \frac{db\,da}{a}, \\ T_f(b,a) &= \langle f, \psi_{(b,a)} \rangle \end{aligned} \tag{4}$$

strongly in $L^2(\mathbb{R})$.

$T_f \in L^2(\mathbb{R} \times \mathbb{R}_+^*)$ is called the wavelet transform of f with respect to the analyzing wavelet ψ. If ψ is sufficiently well localized in time and frequency (i.e., both ψ and $\hat{\psi}$ have sufficient decay at infinity), T_f gives information on the time-frequency localization of f. Conversely, equation (4) states that the wavelet transform is invertible on its range, allowing the reconstruction of the analyzed function from its wavelet transform.

If one restricts oneself to the case of the Hardy space $H^2(\mathbb{R})$, a weaker admissibility condition (concerning only the positive frequency part of ψ) is sufficient. Simply assuming that

$$c_\psi = \int_0^\infty \left|\hat{\psi}(u)\right|^2 \frac{du}{u} = 1, \tag{5}$$

Theorem 2.1 holds for any $f \in H^2(\mathbb{R})$.

Let $\psi(x)$ be an infinitesimal wavelet and let $\Phi(x)$ be such that

$$\left|\hat{\Phi}(\xi)\right|^2 = \int_{|\xi|}^\infty \left|\hat{\psi}(u\,\mathrm{sgn}(\xi))\right|^2 \frac{du}{u}. \tag{6}$$

In other words, $|\hat{\psi}(u\xi)|^2 = -u\,\partial_u|\hat{\Phi}(u\xi)|^2$ for all $\xi \in \mathbb{R}$, and $\lim_{\xi\to\infty} |\hat{\Phi}(\xi)|^2 = 0$. Φ is called a *scaling function*, and one associates the corresponding family of functions with it:

$$\Phi_{(b,a)}(x) = \frac{1}{a}\,\Phi\left(\frac{x-b}{a}\right). \tag{7}$$

Given $f \in L^2(\mathbb{R})$ consider its smoothing (with respect to Φ) at scale a:

$$s_a(x) = \int_{\mathbb{R}} \langle f, \Phi_{(b,a)}\rangle \Phi_{(b,a)}(x)\,db; \tag{8}$$

that is,

$$s_a(x) = \int_a^\infty d_u(x)\,\frac{du}{u}, \tag{9}$$

where $d_a(x)$ stands for the details of $f(x)$ at scale a:

$$d_a(x) = \int_{\mathbb{R}} \langle f, \psi_{(b,a)}\rangle \psi_{(b,a)}(x)db. \tag{10}$$

Then, $s_a \in L^2(\mathbb{R})$ and one has the following decomposition, whose proofs are immediate from that of Theorem 2.1.

COROLLARY 2.2. *Let ψ be an infinitesimal wavelet, and Φ be an associated scaling function. Then any $f \in L^2(\mathbb{R})$ can be expressed as follows:*

$$f = \lim_{a\to 0} s_a = s_{a_0} + \int_0^{a_0} d_a\,\frac{da}{a} \tag{11}$$

(for any $a_0 \in \mathbb{R}_+^*$) strongly in $L^2(\mathbb{R})$.

The corollary also holds in the $H^2(\mathbb{R})$ context. Let us now set

$$D_j(x) = \int_{2^{-j-1}}^{2^{-j}} d_a(x)\,\frac{da}{a}\,. \tag{12}$$

Then, $D_j \in L^2(\mathbb{R})$ represents the details of $f(x)$ visible at scale 2^{j+1}, not at scale 2^j, and

$$\widehat{D_j}(\xi) = \hat{f}(\xi) \int_{2^{-j-1}}^{2^{-j}} \left|\hat{\psi}(a\xi)\right|^2 \frac{da}{a} \text{ a.e.} \tag{13}$$

Introducing the function $\Psi(x)$ such that

$$\left|\widehat{\Psi}(\xi)\right|^2 = \int_{\frac{1}{2}}^{1} \left|\hat{\psi}(a\xi)\right|^2 \frac{da}{a}, \tag{14}$$

one has

$$\widehat{D_j}(\xi) = \hat{f}(\xi) \left|\widehat{\Psi}(2^{-j}\xi)\right|^2. \tag{15}$$

We will refer to the $\Psi(x)$ function as the (*global or integrated*) *wavelet*. Note that equation (14) does not completely define the wavelet Ψ. Once again, one can restrict oneself to wavelets with positive-valued Fourier transform, but this is not necessary. By construction, the integrated wavelets lead to a partition of unity in the Fourier space as follows:

$$\sum_{j=-\infty}^{+\infty} \left|\widehat{\Psi}(2^j\xi)\right|^2 = \left|\widehat{\Phi}(2^{j_0}\xi)\right|^2 + \sum_{j=-\infty}^{j_0} \left|\widehat{\Psi}(2^j\xi)\right|^2 = 1 \tag{16}$$

for all $\xi \in \mathbb{R}$. We also have

$$|\widehat{\Psi}(\xi)|^2 = |\hat{\Phi}(\xi/2)|^2 - |\hat{\Phi}(\xi)|^2. \tag{17}$$

Defining the dilates and translates of Φ^j and Ψ^j as

$$\begin{aligned} \Phi_b^j(x) &= 2^{-j}\Phi\big(2^{-j}(x-b)\big), \\ \Psi_b^j(x) &= 2^{-j}\Psi\big(2^{-j}(x-b)\big), \end{aligned} \tag{18}$$

we then have the following theorem.

THEOREM 2.3. *Let ψ be an infinitesimal wavelet and Ψ and Φ be associated integrated wavelet and scaling functions as in equations* (14) *and* (6). *Then, any $f \in L^2(\mathbb{R})$ can be decomposed as*

$$f = \int_{\mathbb{R}} \langle f, \Phi_b^{j_0}\rangle \, \Phi_b^{j_0} \, db + \sum_{j=-\infty}^{j_0} \int_{\mathbb{R}} \langle f, \Psi_b^j\rangle \, \Psi_b^j \, db \tag{19}$$

strongly in $L^2(\mathbb{R})$.

We shall use the following notation:

$$\begin{aligned} T_j f(b) &= \langle f, \Psi_b^j\rangle, \\ S_j f(b) &= \langle f, \Phi_b^j\rangle. \end{aligned} \tag{20}$$

2.2. The linear scheme. It is well known that the reconstructing and analyzing wavelets can be decoupled, i.e., one can use different infinitesimal wavelets for the computation of the coefficients and the reconstruction of the analyzed function from the coefficients. In such a case, the admissibility condition (1) has to be modified accordingly.

A particular example of such a decoupling which has been known for a long time consists of formally taking a Dirac distribution for the reconstructing wavelet. Assuming that

$$k_\psi = \int_0^\infty \hat{\psi}(u)\,\frac{du}{u} = \int_0^\infty \hat{\psi}(-u)\,\frac{du}{u} = 1, \tag{21}$$

instead of equation (1), one has the following decomposition of any $f \in L^2(\mathbb{R})$:

$$f(x) = \int_{\mathbb{R}_+^*} \langle f, \psi_{(x,a)} \rangle\,\frac{da}{a} \tag{22}$$

strongly in $L^2(\mathbb{R})$. This is the so-called Morlet reconstruction formula of f from its wavelet coefficients. Such a linear analysis (linear in the ψ function) generates a continuous multiresolution analysis as follows: Introduce the *linear scaling function* $\varphi \in L^1(\mathbb{R})$ defined by

$$\hat{\varphi}(\xi) = \int_{|\xi|}^\infty \hat{\psi}(u\,\mathrm{sgn}(\xi))\,\frac{du}{u}\ . \tag{23}$$

φ is also such that $\hat{\psi}(u\xi) = -u\,\partial_u \hat{\varphi}(u\xi)$ for all $\xi \in \mathbb{R}$.

Associate with φ the following family of functions:

$$\varphi_{(b,a)}(x) = \frac{1}{a}\,\varphi\left(\frac{x-b}{a}\right). \tag{24}$$

Finally, introduce

$$\delta_a(x) = T_f(x,a) = \langle f, \psi_{(x,a)} \rangle \tag{25}$$

and

$$\sigma_a(x) = \langle f, \varphi_{(x,a)} \rangle\ . \tag{26}$$

One then has the linear analogue of Theorem 2.1 and the corresponding corollary.

THEOREM 2.4. *Let $\psi \in L^1(\mathbb{R})$ be a mother wavelet such that equation* (21) *holds, and let φ be the associated linear scaling function. Then any $f \in L^2(\mathbb{R})$ can be decomposed as*

$$\begin{aligned} f &= \lim_{a\to 0} \sigma_a \\ &= \sigma_{a_0} + \int_0^{a_0} \delta_a\,\frac{da}{a},\ a_0 \in \mathbb{R}_+^* \\ &= \int_0^\infty \delta_a\,\frac{da}{a} \end{aligned} \tag{27}$$

strongly in $L^2(\mathbb{R})$.

The integrated wavelets are then defined as

$$\widehat{\Theta}(\xi) = \int_{\frac{1}{2}}^{1} \hat{\psi}(a\xi)\,\frac{da}{a} \tag{28}$$

and yield a partition of unity in the Fourier space

$$\sum_{j=-\infty}^{+\infty} \widehat{\Theta}(2^j\xi) = \hat{\varphi}(2^{j_0}\xi) + \sum_{j\le j_0} \widehat{\Theta}(2^j\xi) = 1. \tag{29}$$

The linear wavelets still appear as differences of smoothings at two consecutive scales as follows:

$$\widehat{\Theta}(\xi) = \hat{\varphi}(\xi/2) - \hat{\varphi}(\xi), \tag{30}$$

and every $f \in L^2(I\!R)$ decomposes as

$$f(x) = \sum_{j=-\infty}^{\infty} \langle f, \Theta_x^j\rangle = \langle f, \varphi_x^{j_0}\rangle + \sum_{j\le j_0} \langle f, \Theta_x^j\rangle, \tag{31}$$

where $\Theta_b^j(x) = \Theta^j(x-b)$.

3. Associated approximate filters.

3.1. Pyramidal algorithms. Let us recall the usual algorithmic structure associated with a multiresolution analysis. Let ϕ and ψ be, respectively, a scaling function and a wavelet associated with the multiresolution analysis, and set $\forall f \in L^2(I\!R)$,

$$\begin{aligned} T_j f(n) &= \langle f, \psi_{jn}\rangle = 2^{-j}\int_{I\!R} f(x)\psi(2^{-j}(x-n))^* dx, \\ S_j f(n) &= \langle f, \phi_{jn}\rangle = 2^{-j}\int_{I\!R} f(x)\phi(2^{-j}(x-n))^* dx. \end{aligned} \tag{32}$$

Note that for any value of the scale parameter $a = 2^j$, we sample the corresponding wavelet and scaling function transform at unit sampling frequency.

Then, if ψ and ϕ are related by

$$\begin{aligned} \hat{\phi}(2\xi) &= m_0(\xi)\hat{\phi}(\xi), \\ \hat{\psi}(2\xi) &= m_1(\xi)\hat{\phi}(\xi), \end{aligned} \tag{33}$$

where m_0 and m_1 are the 2π-periodic low-pass and high-pass filters

$$\begin{aligned} m_0(\xi) &= \sum_k h_k e^{ik\xi}, \\ m_1(\xi) &= \sum_k g_k e^{ik\xi}, \end{aligned} \tag{34}$$

then the coefficients may be computed using the following pyramidal algorithm

$$\begin{aligned} T_j f(n) &= \sum_k g_k^* S_{j-1} f(n - k2^{-j-1}), \\ S_j f(n) &= \sum_k h_k^* S_{j-1} f(n - k2^{-j-1}). \end{aligned} \tag{35}$$

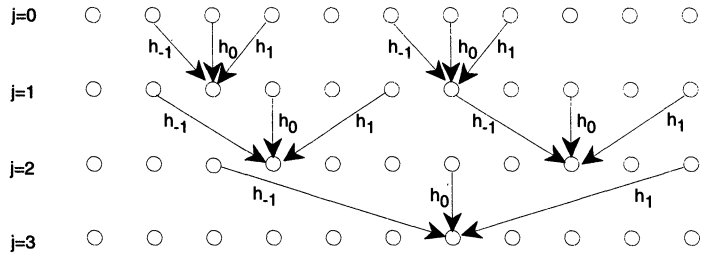


FIG. 1. *QMF algorithm associated with the wavelet transform on a fine grid.*

The algorithm is called pyramidal since scaled copies of the same filters are used throughout the calculation and the coefficients are obtained by successive convolutions with such filters. It is easy to see that the total number of multiplications necessary to process N samples of, say, $S_0 f$ is proportional to $N \log(N)$. It is schematically described in Fig. 1 (in the particular situation where the $m_0(\xi)$ filter has only three nonvanishing coefficients).

3.2. Approximate filters. Now we address the problem of discretization of the previous wavelet decompositions. Up to now we have only obtained decompositions that are discrete with respect to the scale and continuous with respect to the position. The problem is that no discrete filters are "a priori" available.

We shall work with both the linear and the bilinear analysis-reconstruction schemes at the beginning and specify our choice later on. From now on we shall assume that a pair of functions

$$m_0(\xi) = \frac{\widehat{\Phi}(2\xi)}{\widehat{\Phi}(\xi)}, \tag{36}$$

$$m_1(\xi) = \frac{\widehat{\Psi}(2\xi)}{\widehat{\Phi}(\xi)} \tag{37}$$

can be defined almost everywhere in $\mathbb{R}$. This is clearly the case in the bilinear scheme, where $|\widehat{\Phi}|$ is monotonic for both $\xi \geq 0$ and $\xi \leq 0$. The problem is that, in general, such an m_0-function is not 2π-periodic and thus cannot be used in a pyramidal algorithm.

Nevertheless, a modification is possible. Indeed, if $\widehat{\Phi}(\xi)$ is "concentrated" around the origin in, say, the interval $[-\pi, \pi]$,[1] then one may expect that $\widehat{\Phi}$ "does not see too much" the nonperiodicity of $m_0(\xi)$, and that $m_0(\xi)\widehat{\Phi}(\xi)$ can be well approximated by $m_0^a(\xi)\widehat{\Phi}(\xi)$ for some 2π-periodic function $m_0^a(\xi)$.[2]

Here let us introduce for convenience the following subspace of $L^2(\mathbb{R})$:

$$\mathcal{U}_0 = \left\{ f \in L^2(\mathbb{R}), f = \sum \alpha_k \Phi(x-k), \{\alpha_k\} \in \ell^2(\mathbb{Z}) \right\}. \tag{38}$$

[1] This assumption is motivated by the fact that we will sample the wavelet transform and the scaling function transform at unit sampling frequency.

[2] A natural candidate for $m_0^a(\xi)$ is the periodization $\sum_k m_0(\xi + 2\pi k)$ of $m_0(\xi)$, but as we shall see, there are many other choices.

We will assume that the collection $\{\Phi(x-k),\, k\in\mathbb{Z}\}$ is a Riesz basis of $\mathcal{U}_0$ or, equivalently, that there exist two finite and nonzero constants A and B such that

$$A \le \sum_k |\widehat{\Phi}(\xi+2\pi k)|^2 \le B \text{ a.e.} \tag{39}$$

Then, it follows from general results that there exists a function $\chi\in L^2(\mathbb{R})$ such that the sequence $\{\chi(x-k),\, k\in\mathbb{Z}\}$ is the basis of $\mathcal{U}_0$ biorthogonal to $\{\Phi(x-k),\, k\in\mathbb{Z}\}$. χ is given by its Fourier transform

$$\widehat{\chi}(\xi) = \frac{\widehat{\Phi}(\xi)}{\sum_k |\widehat{\Phi}(\xi+2\pi k)|^2}. \tag{40}$$

Now consider the discretization of the functions $T_j f(x)$ and $S_j f(x)$ that we denote by $T_j^d f(n)$ and $S_j^d f(n)$, respectively as follows:

$$\begin{aligned} S_j^d f(n) &= S_j f(n)\ \forall n\in\mathbb{Z}, \\ T_j^d f(n) &= T_j f(n)\ \forall n\in\mathbb{Z}. \end{aligned} \tag{41}$$

Let $m_0^a(\xi)\in L^2([-\pi,\pi])$ and $m_1^a(\xi)\in L^2([-\pi,\pi])$ be two (2π-periodic) candidates for approximate filters and denote by $\{h_k^a, k\in\mathbb{Z}\}$ and $\{g_k^a, k\in\mathbb{Z}\}$ their respective Fourier coefficients. Then, we will set

$$\begin{cases} T_1^a f(n) &= \sum_k g_k^{a*} S_0^d f(n-k), \\ S_1^a f(n) &= \sum_k h_k^{a*} S_0^d f(n-k), \end{cases} \tag{42}$$

and for $j>1$,

$$\begin{cases} T_j^a f(n) &= \sum_k g_k^{a*} S_{j-1}^a f(n-2^{j-1}k), \\ S_j^a f(n) &= \sum_k h_k^{a*} S_{j-1}^a f(n-2^{j-1}k). \end{cases} \tag{43}$$

Our purpose is to compare such "algorithmic expressions"[3] with the exact expressions $S_j^d f$ and $T_j^d f$, and find "best approximants" for the m_0^a and m_1^a filters.

The first remark is the following proposition.

PROPOSITION 3.1.

1. $\mathrm{Ker}(S_0^d) = \mathcal{U}_0^\perp$.
2. $S_j^a\cdot\mathcal{U}_0^\perp = T_j^a\cdot\mathcal{U}_0^\perp = 0$ *for any* $j=1,\dots$.

Proof. The first part is a direct consequence of the definition of $\mathcal{U}_0$ and implies the second part by definition (43). □

Our main result is the following theorem.

THEOREM 3.2. *Let $\Phi(x)$ and $\Psi(x)$ be the scaling function and the integrated wavelet, respectively, associated with the infinitesimal wavelet $\psi(x)$, and let $m_0(\xi)$ and $m_1(\xi)$ be the associated low-pass and high-pass filters. For $i=0,1$ set*

$$\mu(m_i, m_i^a) = \left[\int_{\mathbb{R}} \left|(m_i^a(\xi)-m_i(\xi))\widehat{\Phi}(\xi)\right|^2 d\xi\right]^{1/2}. \tag{44}$$

Then the following properties are satisfied:

[3] We do this because this is precisely what is numerically computed in practice.

1. *There exists a unique pair of* 2π*-periodic filters* $m_i^a(\xi) = \tilde{m}_i(\xi)$ *minimizing* $\mu(m_i, m_i^a)$ *given by*

$$\tilde{m}_0(\xi) = \frac{\sum_{k\in\mathbf{Z}} \hat{\Phi}(\xi+2\pi k)^* \hat{\Phi}(2(\xi+2\pi k))}{\sum_{k\in\mathbf{Z}} |\hat{\Phi}(\xi+2\pi k)|^2}, \tag{45}$$

$$\tilde{m}_1(\xi) = \frac{\sum_{k\in\mathbf{Z}} \hat{\Phi}(\xi+2\pi k)^* \hat{\Psi}(2(\xi+2\pi k))}{\sum_{k\in\mathbf{Z}} |\hat{\Phi}(\xi+2\pi k)|^2}. \tag{46}$$

2. *For the above choice of filters, and setting*

$$C_i = \operatorname*{ess\,sup}_{\xi\in\mathbb{R}} |\tilde{m}_i(\xi)|, \quad i = 0, 1, \tag{47}$$

the following inequalities hold:

$$\|S_j^a f - S_j^d f\|_\infty \le \mu(\tilde{m}_0, m_0) 2^{(1-j)/2} \frac{1-(C_0\sqrt{2})^j}{1-C_0\sqrt{2}} \|f\|_2, \tag{48}$$

(49)

$$\|T_j^a f - T_j^d f\|_\infty \le 2^{(1-j)/2} \left(\mu(\tilde{m}_1, m_1) + C_1 \mu(\tilde{m}_0, m_0) \sqrt{2} \frac{1-(C_0\sqrt{2})^{j-1}}{1-C_0\sqrt{2}} \right) \|f\|_2.$$

3. *For any* $f \in \mathcal{U}_0$,

$$\begin{aligned} S_1^a f &= S_1^d f, \\ T_1^a f &= T_1^d f. \end{aligned} \tag{50}$$

Before giving the proof of the theorem, let us give the following immediate corollary.

COROLLARY 3.3. *Assume that the infinitesimal wavelet* $\psi(x)$ *is associated with a linear analysis-reconstruction scheme. Then the associated approximate quadrature mirror filters* (*QMFs*) $\tilde{m}_i(\xi)$ *given by equations* (45) *and* (46) *satisfy*

$$\tilde{m}_0(\xi) + \tilde{m}_1(\xi) = 1, \tag{51}$$

so that we have the reconstruction algorithm

$$S_0 f(n) = \sum_{j\ge 0} T_j^a(n). \tag{52}$$

This reconstruction formula is the discrete counterpart of Morlet's reconstruction formula (22). It was also obtained in a slightly different context by Saito and Beylkin [5].

Proof of the theorem. Using the inequality $\|f\|_\infty \le \|\hat{f}\|_1/2\pi$, we shall work directly in the Fourier space. First of all, we clearly have

$$\begin{aligned} \|\widehat{S_1^a f} - \widehat{S_1^d f}\|_1 &\le \sum_{k\in\mathbf{Z}} \int_0^{2\pi} \left| [m_0^a(\xi) - m_0(\xi+2\pi k)] \hat{\Phi}(\xi+2\pi k) \hat{f}(\xi+2\pi k) \right| d\xi \\ &\le \int_{\mathbb{R}} \left| [m_0^a(\xi) - m_0(\xi)] \hat{\Phi}(\xi) \hat{f}(\xi) \right| d\xi \\ &\le 2\pi \|f\|_2 \left[\int_{\mathbb{R}} [m_0^a(\xi)\hat{\Phi}(\xi) - \hat{\Phi}(2\xi)|^2 d\xi \right]^{1/2} = 2\pi \mu(m_0^a, m_0) \|f\|_2 \end{aligned} \tag{53}$$

(the last inequality comes from the Cauchy–Schwartz inequality). This explains the occurence of such a term in our formulation. The minimization of this term is a classical problem and leads to

$$\int_{\mathbb{R}} \Phi(x+k)^* \left[\sum_l h_l^a \Phi(x+l) - \frac{1}{2}\Phi\left(\frac{x}{2}\right) \right] dx = 0 \quad \forall k \in \mathbb{Z} \tag{54}$$

or, otherwise stated,

$$\int_0^{2\pi} e^{ik\xi} \left[m_0^a(\xi) \sum_{l\in\mathbb{Z}} |\widehat{\Phi}(\xi+2\pi l)|^2 - \sum_{l\in\mathbb{Z}} \widehat{\Phi}(\xi+2\pi l)^* \widehat{\Phi}\big(2(\xi+2\pi l)\big) \right] d\xi = 0 \tag{55}$$
$$\forall k \in \mathbb{Z}.$$

The unique solution is precisely that given in equation (45). The estimation of $||T_1^a f - T_1^d f||_\infty$ is completely similar and leads to the approximate filter given in equation (45). The details are left to the reader.

Let us now consider larger scales. Before going into the details, let us introduce for convenience the following "intermediate" sequences:

$$\begin{cases} T_j^i f(n) &= \sum_k g_k^{a*} S_{j-1}^d f(n - 2^{j-1}k), \\ S_j^i f(n) &= \sum_k h_k^{a*} S_{j-1}^d f(n - 2^{j-1}k). \end{cases} \tag{56}$$

Then, clearly,

$$\begin{aligned} ||\widehat{S_j^a f} - \widehat{S_j^d f}||_1 &\le ||\widehat{S_j^a f} - \widehat{S_j^i f}||_1 + ||\widehat{S_j^i f} - \widehat{S_j^d f}||_1, \\ ||\widehat{T_j^a f} - \widehat{T_j^d f}||_1 &\le ||\widehat{T_j^a f} - \widehat{T_j^i f}||_1 + ||\widehat{T_j^i f} - \widehat{T_j^d f}||_1. \end{aligned} \tag{57}$$

Again we focus on the approximations $S_j^a f$, the proof for the details $T_j^a f$ being completely similar.

(58)

$$\begin{aligned} ||\widehat{S_j^i f} - \widehat{S_j^d f}||_1 &= \int_0^{2\pi} \Big| \sum_{k\in Z} m_0\big(2^{j-1}(\xi+2\pi k)\big) \widehat{\Phi}\big(2^{j-1}(\xi+2\pi k)\big) \hat{f}(\xi+2\pi k) \\ &\qquad - \sum_{k\in\mathbb{Z}} m_0^a\big(2^{j-1}\xi\big) \widehat{\Phi}\big(2^{j-1}(\xi+2\pi k)\big) \hat{f}(\xi+2\pi k) \Big| d\xi \\ &\le 2\pi ||f||_2 \left[\int_{\mathbb{R}} \left| m_0^a\big(2^{j-1}\xi\big) \widehat{\Phi}\big(2^{j-1}\xi\big) - \widehat{\Phi}\big(2^j\xi\big) \right|^2 d\xi \right]^{1/2} \\ &\le 2\pi ||f||_2 2^{(1-j)/2} \left[\int_{\mathbb{R}} \left| m_0^a(\xi) \widehat{\Phi}(\xi) - \widehat{\Phi}(2\xi) \right|^2 d\xi \right]^{1/2} \\ &= 2\pi 2^{(1-j)/2} \mu(m_0^a, m_0) ||f||_2. \end{aligned}$$

The second term is estimated as follows:

$$\begin{aligned} ||\widehat{S_j^a f} - \widehat{S_j^i f}||_1 &\le \textstyle\int_{\mathbb{R}} \big| m_0^a\big(2^{j-1}\xi\big) \big| \big| \widehat{S_{j-1}^a f}(\xi) - \widehat{S_{j-1}^d f}(\xi) \big| d\xi \\ &\le \operatorname{ess\,sup}_{\xi\in\mathbb{R}} |m_0^a| ||\widehat{S_{j-1}^a f} - \widehat{S_{j-1}^d f}||_1. \end{aligned} \tag{59}$$

Summarizing, for $m_0^a(\xi) = \tilde{m}_0(\xi)$ we have

(60)

$$\begin{aligned}\|\widehat{S_j^a f} - \widehat{S_j^d f}\|_1 &\le 2\pi 2^{(1-j)/2}\mu(\tilde{m}_0, m_0)\|f\|_2 + C_0\|\widehat{S_{j-1}^a f} - \widehat{S_{j-1}^d f}\|_1 \\ &\le 2\pi 2^{(1-j)/2}\mu(\tilde{m}_0, m_0)\left(1 + C_0\sqrt{2} + (C_0\sqrt{2})^2 + \cdots + (C_0\sqrt{2})^{j-1}\right)\|f\|_2. \\ &\le 2\pi 2^{(1-j)/2}\mu(\tilde{m}_0, m_0)\frac{1-(C_0\sqrt{2})^j}{1-C_0\sqrt{2}}\|f\|_2.\end{aligned}$$

The same kind of estimate yields the error estimate for the $T_j^a f$ coefficients. This achieves the proof of the two first items of the theorem.

Let us turn to the third part of the theorem. Then, let us assume that $f \in \mathcal{U}_0$. This means that in the Fourier space, f is of the form

$$\hat{f}(\xi) = F(\xi)\widehat{\Phi}(\xi) \tag{61}$$

for some 2π-periodic function $F \in L^2([-\pi, \pi])$. Then, an explicit computation of $\widehat{S_1^a f} - \widehat{S_1^d f}$ yields

$$\begin{aligned}\widehat{S_1^a f} - \widehat{S_1^d f} &= \sum_{k\in\mathbb{Z}} \left(m_0^a(\xi)^* - m_0(\xi + 2\pi k)^*\right)\left|\widehat{\Phi}(\xi + 2\pi k)\right|^2 F(\xi) \\ &= 0 \text{ if } m_0^a = \tilde{m}_0.\end{aligned} \tag{62}$$

This concludes the proof of the theorem. □

Remark (asymptotic behaviour). It is interesting to analyze the asymptotic behaviour of the estimates when $j \to \infty$. Consider, for instance, the estimate of $\|S_j^a f - S_j^d f\|_\infty$; the coefficient of $\|f\|_2$ is

$$\|S_j^a f - S_j^d f\|_\infty \sim_{j\to\infty} \mu(\tilde{m}_0, m_0)\sqrt{2}C_0^j/(C_0\sqrt{2} - 1)\,\|f\|_2 \tag{63}$$

(for $C_0\sqrt{2} \neq 1$) in the limit. The limit is finite for $C_0 = 1$ and zero for $C_0 < 1$, while it diverges for $C_0 > 1$. In the two first cases, this means that the accumulation of errors resulting from the approximate algorithm is compensated by the fact that $S_j f$, lying at larger and larger scales, is sampled at the same frequency all the time. This shows that "redundancy implies stability".

3.3. Decay of approximate filter coefficients. The localization properties of the $\{h_k\}$ (and thus $\{g_k\}$) approximate filters can be directly related to the regularity properties of the scaling function as follows.

THEOREM 3.4. *Let $\phi \in L^1(\mathbb{R})$ be a p-times differentiable scaling function and $\tilde{m}_0(\xi) = \sum_k h_k e^{ik\xi}$ be the low-pass filter defined according to equation* (45). *Then, if for any $m = 0, 1, \ldots, p$,*

$$\left|\frac{d^m\hat{\phi}(\xi)}{d\xi^m}\right| \le \frac{K_m}{(1+|\xi|)^{1/2+\epsilon}} \tag{64}$$

for some positive constants K_m, ϵ, then

$$h_k = O\left(k^{-p}\right). \tag{65}$$

Proof. Assume that $\hat{\phi}(\xi)$ is p-times differentiable. Then, after p differentiations, equation (45) yields

$$\frac{d^p \tilde{m}_0(\xi)}{d\xi^p} = \frac{\sum_k G(\xi + 2\pi k)}{\left[\sum_k |\hat{\phi}(\xi + 2\pi k)|^2\right]^p}, \tag{66}$$

where $G(\xi)$ is a finite linear combination of terms of the form

$$\frac{d^m \hat{\phi}(\xi)}{d\xi^m} \frac{d^{p-m} \hat{\phi}(\xi)^*}{d\xi^{p-m}}$$

and

$$\frac{d^m \hat{\phi}(\xi)}{d\xi^m} \frac{d^{p-m} \hat{\phi}(2\xi)^*}{d\xi^{p-m}}$$

and their complex conjugates. Then, estimate (64) gives

$$\frac{d^p \tilde{m}_0(\xi)}{d\xi^p} \in L^\infty([0, 2\pi]) \tag{67}$$

and

$$\frac{d^p \tilde{m}_0(\xi)}{d\xi^p} \in C(I\!R). \tag{68}$$

Moreover,

$$k^p h_k = \frac{i^{-p}}{2\pi} \int_0^{2\pi} \frac{d^p \tilde{m}_0(\xi)}{d\xi^p} e^{-ik\xi} d\xi \tag{69}$$

leads to

$$|k|^p |h_k| \leq \frac{1}{2\pi} \int \left| \frac{d^p \tilde{m}_0(\xi)}{d\xi^p} \right| d\xi \leq \left\| \frac{d^p \tilde{m}_0}{d\xi^p} \right\|_\infty, \tag{70}$$

which proves the theorem. □

Thus, under some weak assumptions on the scaling function, it is possible to get well-localized filters. However, this problem is completely independent of the accuracy problem addressed in the previous section.

Note also that Theorem 3.4 should be compared with similar results in the case of classical multiresolution analysis, which leads to the notion of r-regular multiresolution analysis (see [6]).

3.4. The bilinear scheme. We have seen in the corollary of the previous section that the approximate filters $\tilde{m}_0$ and $\tilde{m}_1$ given in (45) and (46) are ideally adapted to the linear analysis-reconstruction scheme. However, in the bilinear case, $\tilde{m}_0$ and $\tilde{m}_1$ cannot be directly used to reconstruct the analyzed function from the approximate coefficients, since they do not fulfill the QMF condition

$$|\tilde{m}_0(\xi)|^2 + |\tilde{m}_1(\xi)|^2 \neq 1 \text{ in general.} \tag{71}$$

Then, a possibility is to use different filters for the reconstruction, for instance, use $\tilde{m}_0$ as the low-pass filter and

$$\tilde{m}_1^r(\xi) = \frac{1 - |\tilde{m}_0(\xi)|^2}{\tilde{m}_1(\xi)} \tag{72}$$

as the high-pass filter. In such a case one must be careful with the zeros of the $\tilde{m}_1(\xi)$ filter.

As an alternative, the same kind of analysis as before can be performed in the bilinear analysis-reconstruction scheme. The previous arguments must be applied to the details and approximations instead of the wavelet coefficients themselves:

$$\widehat{s_j}(\xi) = \left|\widehat{\Phi}(2^j\xi)\right|^2 \hat{f}(\xi). \tag{73}$$

Again using approximate filters to evaluate the coefficients, one is naturally led to the quantity

$$\widehat{s_j^a}(\xi) = |m_0^a(\xi)|^2 \widehat{s_{j-1}^a f}(\xi). \tag{74}$$

At the first step, for instance, one has to evaluate

$$\|s_1^a - s_1^d\|_\infty \leq \|f\|_2 \int_{\mathbb{R}} \left| \left[|m_0^a(\xi)|^2 - |m_0(\xi)|^2 |\widehat{\Phi}(\xi)|^2\right] \right|^2 d\xi. \tag{75}$$

The minimization of such a quantity naturally leads to

$$|m_0^a(\xi)|^2 = \frac{\sum_{k\in\mathbb{Z}} |\widehat{\Phi}(\xi + 2\pi k)|^2 |\widehat{\Phi}(2\xi + 4\pi k)|^2}{\sum_{k\in\mathbb{Z}} |\widehat{\Phi}(\xi + 2\pi k)|^4} \tag{76}$$

and, similarly,

$$|m_1^a(\xi)|^2 = \frac{\sum_{k\in\mathbb{Z}} |\widehat{\Phi}(\xi + 2\pi k)|^2 |\widehat{\Psi}(2\xi + 4\pi k)|^2}{\sum_{k\in\mathbb{Z}} |\widehat{\Phi}(\xi + 2\pi k)|^4}. \tag{77}$$

It is worth noting that in such a case, the bilinear scheme is well suited for this pair of filters and ensures the validity of the usual QMF relation

$$|m_0^a(\xi)|^2 + |m_1^a(\xi)|^2 = 1. \tag{78}$$

Moreover, it is easy to derive the "bilinear counterpart" of Theorem 3.4, relating the length of the approximate filters to the regularity of the scaling function.

3.5. Some complementary remarks. 1. The algorithm described above is actually adapted to the problem of finding approximate discretization of Littlewood–Paley decompositions and is a priori independent of the linear or bilinear schemes derived from continuous wavelet decompositions. In other words, there is no connection between the b discretization problem and the scale discretization (which is not a true discretization in the method reported in §2). Corollary 3.3 simply states that if one considers the filters $\tilde{m}_0$ and $\tilde{m}_1$, the choice of the linear scheme yields simpler reconstruction formulas.

2. Throughout this paper, we have implicitly fixed a reference scale by the choice of a sampling frequency equal to one for all the voices of the wavelet transform. A change of this sampling frequency is equivalent to a global scaling of $L^2(\mathbb{R})$.

3. Assume that we are in the case of a scaling function with exponential decay in the Fourier space (i.e., $\widehat{\Phi}(\xi) \leq C_\phi e^{-\alpha|\xi|}$ for some positive α). Then, it is not very difficult to show that (in the case of a unit sampling frequency) the approximation of the filters obtained by sampling the inverse Fourier transform of m_0 leads to an error

on the scaling function coefficients of the order $e^{-\alpha\pi}$. In the same way, defining the approximate 2π-periodic high-pass filter by the QMF relation leads to the same kind of error estimate for the wavelet coefficients.

4. Obviously, it follows from the expressions of the approximate filters (both in the linear and bilinear schemes) that if Φ and Ψ are associated with a usual multiresolution analysis, with 2π-periodic filters, one recovers $m_0^a = m_0$ and $m_1^a = m_1$.

5. It was shown in [2] how to use Calderón's formula to get descriptions of the Fourier space different from the Littlewood–Paley one by replacing the powers of 2 by an arbitrary monotonic sequence of scale parameters. It sounds reasonable to think of corresponding approximate algorithms similar to the one described above, at least for rational scale parameters. However, this has not been done at the present time.

4. Examples. There are many examples of continuous wavelets for which an efficient algorithm is needed. Here we describe some very simple examples (the filter coefficients have been computed using the Mathematica package).

4.1. The LOG and DOG wavelets. The LOG wavelets are widely used in the context of computer vision. LOG stands for Laplacian of Gaussians. As stressed in [2], in the linear decomposition-reconstruction scheme, if

$$\psi(x) = \frac{1}{\sqrt{\pi}}(1-x^2)e^{-x^2/2}, \tag{79}$$

the associated scaling function and integrated wavelet are given by

$$\varphi(x) = \frac{1}{2\sqrt{2\pi}}e^{-x^2/2} \tag{80}$$

and

$$\Theta(x) = \frac{1}{2\sqrt{2\pi}}\left(e^{-x^2/2} - 2e^{-2x^2}\right). \tag{81}$$

The integrated wavelet is then a DOG (Difference of Gaussians) wavelet and it is no problem to derive the detail coefficients $T_j^d f$ from the approximations $S_j^d f$. But one clearly needs an efficient algorithm to compute the approximations. In general, the scaling function and wavelet have to be scaled properly for the corresponding transforms to be accurately sampled at unit sampling frequency. We shall then consider more general scaling functions

$$\hat{\varphi}(\xi) = e^{-\xi^2/\alpha} \tag{82}$$

with the corresponding integrated wavelets. In Figs. 2 and 3, we give as examples the plots of the approximate low-pass filters $\tilde{m}_0(\xi)$ (the high-pass filter $\tilde{m}_1(\xi)$ is easy to deduce), and the coefficients of $\tilde{m}_0(\xi)$ and $\tilde{m}_1(\xi)$ for $\alpha = 4$ and $\alpha = 6$, respectively. It is worth noting that in both cases (and, in fact, for any positive α) the $\{h_k\}$ and $\{g_k\}$ sequences are rapidly decreasing as a consequence of Theorem 3.4.

4.1.1. $\alpha = 4$. The h_k coefficients are

$$\begin{aligned}
&\{0.3256327400276189, 0.23348983204217, 0.085798244731082,\\
&\ 0.01625749135447543, 0.0015489926732795, 0.000092240628493268 5,\\
&\ -6.3294\,10^{-6}, 5.04664\,10^{-6}, -3.0372\,10^{-6},\\
&\ 1.84053\,10^{-6}, -1.11595\,10^{-6}, 6.76772\,10^{-7},\\
&\ -4.10463\,10^{-7}, 2.48954\,10^{-7}, -1.50997\,10^{-7}, 9.15844\,10^{-8}\},
\end{aligned} \tag{83}$$

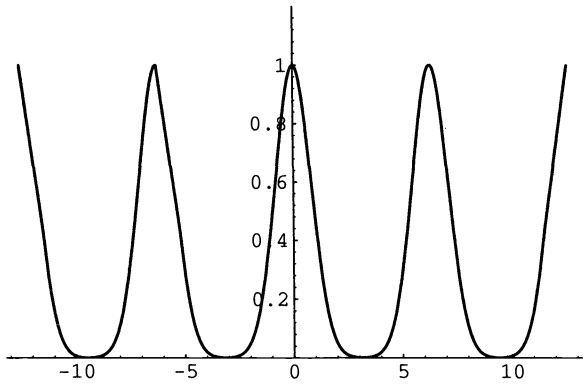


FIG. 2. *Approximate low-pass filter for the DOG wavelet with* $\alpha = 4$.

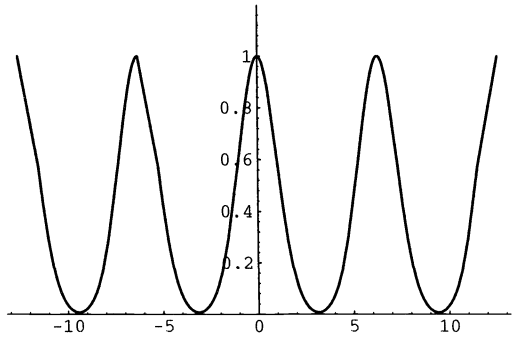


FIG. 3. *Approximate low-pass filter for the DOG wavelet with* $\alpha = 6$.

and the g_k coefficients are

$$
\begin{aligned}
&\{0.6743672599723812, -0.2334898320421699, -0.085798244731082,\\
&-0.01625749135447542, -0.00154899267327945, -0.000092240628493343,\\
&6.3294\,10^{-6}, -5.04664\,10^{-6}, 3.0372\,10^{-6},\\
&-1.84053\,10^{-6}, 1.11595\,10^{-6}, -6.76772\,10^{-7},\\
&4.10463\,10^{-7}, -2.48954\,10^{-7}, 1.50997\,10^{-7}, -9.15844\,10^{-8}\}.
\end{aligned} \tag{84}
$$

4.1.2. $\alpha = 6$. The corresponding h_k coefficients are

$$
\begin{aligned}
&\{0.3972771041574456, 0.2433136948393599, 0.05323383400865474,\\
&0.004788125238023543, -0.0000293448591726243 8,\\
&0.00007751303939803943, -0.00003576989742401233,\\
&0.00001688424764446744, -7.97394\,10^{-6},\\
&3.76645\,10^{-6}, -1.77913\,10^{-6}, 8.40398\,10^{-7},\\
&-3.96976\,10^{-7}, 1.87518\,10^{-7}, -8.85773\,10^{-8}, 4.18409\,10^{-8}\},
\end{aligned} \tag{85}
$$

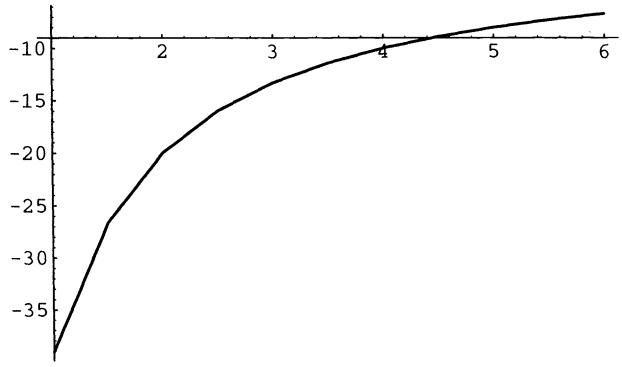


FIG. 4. *Logarithm of* $\mu(m_0, \tilde{m}_0)$ *as a function of* α.

and the g_k coefficients are

$$
\begin{aligned}
&\{0.6027228958425545, -0.2433136948393598, -0.05323383400865478,\\
&\ -0.004788125238023527, 0.0000293448591726369 5,\\
&\ -0.0000775130393980 8579, 0.0000357698974239 5847,\\
&\ -0.0000168842476444 5047, 7.97394\,10^{-6},\\
&\ -3.76645\,10^{-6}, 1.77913\,10^{-6}, -8.40398\,10^{-7},\\
&\ 3.96976\,10^{-7}, -1.87518\,10^{-7}, 8.85773\,10^{-8}, -4.18409\,10^{-8}\}.
\end{aligned}
\tag{86}
$$

4.1.3. Precision of the algorithm. As we have seen, the estimate of the accuracy of the approximate algorithm is governed by the functional $\mu(m_0, \tilde{m}_0)$. Here we present numerical estimation of this quantity for the DOG wavelets for various values of the α parameter. For instance, $\mu(m_0, \tilde{m}_0) = 1.03632\ 10^{-17}$ for $\alpha = 1$, $\mu(m_0, \tilde{m}_0) = 1.63909\ 10^{-6}$ for $\alpha = 3$, and $\mu(m_0, \tilde{m}_0) = 0.0012859$ for $\alpha = 6$. Figure 4 represents the logarithm of $\mu(m_0, \tilde{m}_0)$ as a function of α.

4.2. Exponential-type wavelets. These wavelets are real-valued wavelets characterized by their exponential decay in the Fourier space. Let

$$
\hat{\psi}_n(\xi) = \frac{1}{(n-1)!} \frac{|\xi|^n}{\alpha^n} e^{-|\xi|/\alpha}, \tag{87}
$$

$n = 1, \ldots, \infty$ control the number of vanishing moments, and $\hat{\psi}_n(\xi)$ have exponential decay for all n. A direct computation yields the corresponding scaling function

$$
\hat{\varphi}_n(\xi) = e^{-|\xi|/\alpha} \sum_{p=0}^{n-1} \frac{1}{p!} \frac{|\xi|^p}{\alpha^p}. \tag{88}
$$

Note that $\hat{\phi}_n(\xi) \sim_{\xi \sim 0} 1 + O(|\xi|^n)$. Then, $\hat{\phi}_n \in C^{n-1}(I\!R)$ and $|d^m \hat{\phi}/d\xi^m|$ has exponential decay $\forall m = 0, \ldots, n-1$, which implies that $h_k = O(k^{1-n})$.

The integrated wavelets are easy to deduce and the associated low-pass filter is represented in Fig. 5 in the case $n = 1$ (with $\alpha = .3$).

The case $n = 1$ is not very interesting numerically because $\hat{\varphi}$ is nondifferentiable at $\xi = 0$ and the m_0 filter has slow decay. Then we shall show the case $n = 5$, $\alpha = .3$ for which the low-pass filter is shown in Fig. 6.

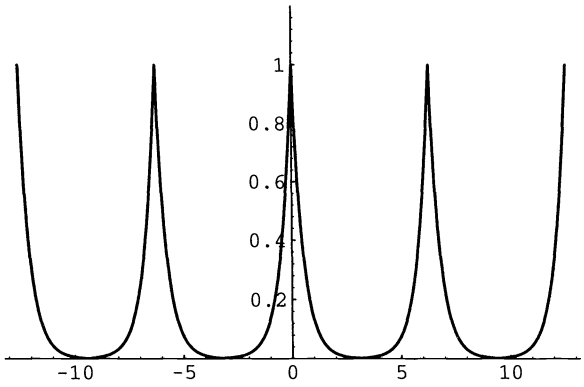


FIG. 5. *Approximate low-pass filter for exponential-type wavelet with* 1 *vanishing moment.*

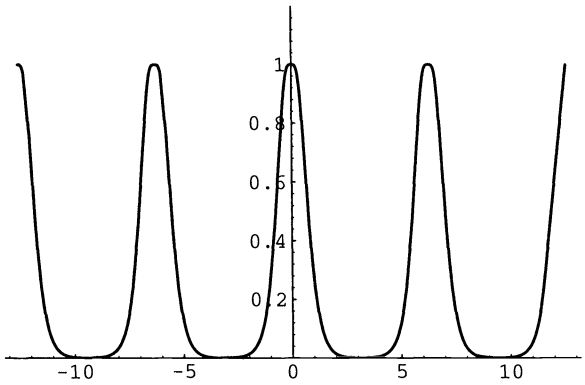


FIG. 6. *Approximate low-pass filter for exponential-type wavelet with* 5 *vanishing moments.*

The 16 top low-pass and high-pass filter coefficients are given by h_k coefficients

$$\begin{aligned}&\{0.2608909, 0.21193501, 0.11802165, 0.047464751,\\ &\ 0.012065044, -0.001261182, -.0.004691885, -0.004500855,\\ &\ -0.0034217616, -0.0023426834, -0.0015318216, -0.00096772754,\\ &\ -0.00060530854, -0.00037287531, -0.00023103465, -0.00014240093\}\end{aligned} \tag{89}$$

and g_k coefficients

$$\begin{aligned}&\{0.7391091, -0.21193501, -0.11802165, -0.047464751,\\ &\ -0.012065044, 0.001261182, 0.004691885, 0.004500855,\\ &\ 0.0034217616, 0.0023426834, 0.0015318216, 0.00096772754,\\ &\ 0.00060530854, 0.00037287531, 0.00023103465, 0.00014240093\}.\end{aligned} \tag{90}$$

It is worth noting that all such coefficients are easy to obtain numerically.

4.3. The Cauchy wavelets. The same filters as before can be used to work with the wavelets that were used by Paul in a quantum mechanical context (they are canonically associated with the radial Schrodinger equation for the hydrogen atom, for instance). They are of the form

$$\hat{\psi}_n(\xi) = \begin{cases} \frac{1}{(n-1)!}\xi^n e^{-\xi} & \text{for positive values of } \xi, \\ 0 & \text{otherwise.} \end{cases} \tag{91}$$

Then $n = 1, \ldots, \infty$ controls the number of vanishing moments and $\hat{\psi}_n(\xi)$ has exponential decay for all n. A direct computation yields the corresponding scaling function

$$\hat{\varphi}_n(\xi) = \begin{cases} e^{-\xi} \sum_{p=0}^{n-1} \frac{1}{p!} \xi^p & \text{for positive values of } \xi, \\ 0 & \text{otherwise.} \end{cases} \tag{92}$$

The first one is particularly interesting since it is canonically related to the Cauchy kernel. Indeed, the scaling function coefficients of a function $f(x) \in H^2(\mathbb{R})$ form an analytic function of $z = b + ia$, that is, the analytic continuation $f(z)$ of $f(x)$ to the upper half-plane. The corresponding wavelet transform is then (up to a factor a) the derivative of $f(z)$ with respect to its imaginary part $T_f(b, a) = -a\partial_a f(b + ia)$.

The previous filters can then be used to get a fast approximate algorithm for wavelet transform with such wavelets. Of course, adapted filters can also be obtained by directly using the formula yielding the $\tilde{m}_i$ filters. However, since $\hat{\phi}_n$ is discontinuous at the origin for any n, such filters are not suitable for numerical use since they have slow decay.

5. Conclusions. In this paper we have described a method that associates a pair of (2π-periodic) filters with a Littlewood–Paley (or dyadic wavelet) decomposition yielding a pyramidal algorithm for the computation of a corresponding approximate transform.

In particular we have shown that in the case where the Littlewood–Paley decomposition comes from a linear scheme of infinitesimal wavelet analysis (as described in [2]), such filters fulfill a kind of linear QMF relation leading to simple reconstruction formulas from the approximate coefficients. Our main result was an estimate of the accuracy of the approximate algorithm. The problem of finding approximate filters was transformed into a minimization problem having a unique solution. Of course, when there already exists a pair of 2π-periodic filters naturally associated with the wavelet, this solution coincides with it.

In the case of the linear scheme of infinitesimal wavelet analysis, we also obtained explicit expressions for approximate filters. It must be noted that in some cases, the error estimates go to zero as the scale becomes larger and larger. This is a result of the fact that the wavelet transform is sampled at a fixed sampling frequency independent of the scale. In such cases, the redundancy of the wavelet transform implies the stability of the algorithm.

As in the case of usual multiresolution analysis, the localization (i.e., decay properties) of the approximate filters is directly related to the regularity of the scaling function.

We also discussed some simple examples, in particular, those of the LOG and DOG wavelets in the linear scheme familiar to computer vision specialists, and wavelets of exponential type. If the corresponding scaling functions are sufficiently well localized in the Fourier space, good error estimates are obtained.

Let us note that n-dimensional generalizations of our method with the tensor-product construction of filters are straightforward.

Acknowledgments. We thank A. Grossmann and Ph. Tchamitchian for stimulating discussions.

REFERENCES

[1] I. DAUBECHIES, *Ten lectures on wavelets*, Society for Industrial and Applied Mathematics, 1992.

[2] M. DUVAL-DESTIN, M. A. MUSCHIETTI, AND B. TORRESANI, *Continuous wavelet decompositions, multiresolution, and contrast analysis*, SIAM J. Math. Anal., 24 (1993), pp. 739–754.
[3] M. FRAZIER, B. JAWERTH, AND G. WEISS, *Littlewood–Paley theory and the study of function spaces*, CBMS-NSF Regional Conf. Ser. in Appl. Math., 1992.
[4] M. HOLSCHNEIDER, R. KRONLAND-MARTINET, J. MORLET, AND PH. TCHAMITCHIAN, *A real-time algorithm for signal analysis with the help of the wavelet transform*, in Wavelets, Time-Frequency Methods and Phase Space, J. M. Combes et al., eds., Inverse Problems and Theoretical Imaging, Springer-Verlag, 1989, pp. 286–297.
[5] N. SAITO AND G. BEYLKIN, *Multiresolution representations using the autocorrelation functions of compactly supported wavelets*, Proc. International Conference on Wavelets and Applications, Y. Meyer and S. Roques, eds., Editions Frontières, Toulouse, 1993.
[6] Y. MEYER, *Ondelettes et opérateurs: I Ondelettes*, Hermann, 1989.
[7] M. SHENSA, *Affine wavelets: Wedding the Atrous and the Mallat algorithms*, IEEE Trans. Signal Process., 40 (1992), pp. 2464–2482.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 944–977, July 1995

008

SEMICLASSICAL ASYMPTOTICS BEYOND ALL ORDERS FOR SIMPLE SCATTERING SYSTEMS *

ALAIN JOYE† AND CHARLES-EDOUARD PFISTER‡

Abstract. The semiclassical limit $\varepsilon \to 0$ of the scattering matrix S associated with the equation $i\varepsilon \frac{d\varphi(t)}{dt} = A(t)\varphi(t)$ is considered. If $A(x)$ is an analytic $n \times n$ matrix whose eigenvalues are real and nondegenerate for all $x \in \mathbf{R}$, the matrix S is computed asymptotically up to errors $O(e^{-\kappa\varepsilon^{-1}}), \kappa > 0$. Moreover, for the case $n = 2$ and under further assumptions on the behavior of the analytic continuations of the eigenvalues of $A(x)$, the exponentially small off-diagonal elements of S are given by an asymptotic expression accurate up to *relative* errors $O(e^{-\kappa\varepsilon^{-1}})$. The adiabatic transition probability for the time-dependent Schrödinger equation, the semiclassical above barrier reflection coefficient for the stationary Schrödinger equation, and the total variation of the adiabatic invariant of a time-dependent classical oscillator are computed asymptotically to illustrate results.

Key words. singular perturbations, turning point theory, semiclassical, and adiabatic approximation, asymptotics of S-matrix

AMS subject classifications. 34E20, 34L25, 81Q20

1. Introduction.

Let us consider the following well-known equations. The first one is the time-dependent Schrödinger equation for a two-level system

$$i\hbar \frac{d\psi(t)}{dt} = H(\varepsilon t)\psi(t) \tag{1.1}$$

$t \in \mathbf{R}, \psi(t) \in \mathcal{H} = \mathbf{C}^2$ and $H(\varepsilon t)$ is a 2×2 self-adjoint linear operator with two distinct real eigenvalues. The parameter ε is positive and small. The second equation is the stationary one-dimensional Schrödinger equation

$$-\hbar^2 \frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = E\psi(x) \tag{1.2}$$

$x \in \mathbf{R}, \psi(x) \in \mathbf{C}$ and $V(x)$ is a bounded real-valued function. The real parameter E is chosen in such a way that

$$E > \sup_{x\in\mathbf{R}} V(x). \tag{1.3}$$

The third equation is the equation of motion of a classical oscillator whose frequency varies with time

$$\ddot{v}(t) = -\omega^2(\varepsilon t)v(t), \quad v(0) = u_0, \quad \dot{v}(0) = u_1. \tag{1.4}$$

This equation is of the same type as (1.2) since we assume that the real-valued function $\omega(t)$ is bounded and such that

$$\inf_{t\in\mathbf{R}} \omega^2(t) > 0. \tag{1.5}$$

* Received by the editors May 23, 1993; accepted for publication (in revised form) January 14, 1994.

† Centre de Physique Théorique, Centre National de la Recherche Scientifique Marseille, Luminy Case 907, F-13288 Marseille Cedex 9, France.

‡ Département de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

For the first two equations we are interested in the behavior of the solution for $t \to +\infty$ or $x \to +\infty$, when the behavior for $t \to -\infty$ or $x \to -\infty$ is fixed. Moreover we want to analyze this scattering situation when ε tends to zero and $\hbar = 1$ for equation (1.1), the so-called adiabatic limit, or $\hbar$ tends to zero for equation (1.2), the so-called semiclassical limit. For the initial value problem (1.4), we consider the adiabatic invariant J defined as twice the ratio of the energy to the frequency

$$J(t,\varepsilon) = \frac{|\dot{v}(t)|^2 + \omega^2(\varepsilon t)|v(t)|^2}{\omega(\varepsilon t)} \tag{1.6}$$

in the limit $\varepsilon \to 0$. More precisely, we are interested in its total variation during the whole evolution

$$\Delta J(\varepsilon) \equiv J(+\infty,\varepsilon) - J(-\infty,\varepsilon).$$

In this respect, we consider (1.4) more as a scattering problem than as an initial value problem. All three problems are very closely related. Let $x = \varepsilon t$ be a rescaled time for equations (1.1) and (1.4). Then equation (1.1) becomes with $\varphi(x) = \psi(t(x))$ and $\hbar = 1$

$$i\varepsilon \frac{d\varphi(x)}{dx} = H(x)\varphi(x). \tag{1.7}$$

On the other hand, defining $u(x) = v(t(x))$ and

$$\varphi(x) = \begin{pmatrix} u(x) \\ i\varepsilon \dfrac{du(x)}{dx} \end{pmatrix}, \tag{1.8}$$

equation (1.4) is equivalent to

$$i\varepsilon \frac{d\varphi(x)}{dx} = \begin{pmatrix} 0 & 1 \\ \omega^2(x) & 0 \end{pmatrix} \varphi(x), \qquad \varphi(0) = \begin{pmatrix} u_0 \\ iu_1 \end{pmatrix}. \tag{1.9}$$

Similarly, with

$$\varphi(x) = \begin{pmatrix} \psi(x) \\ i\varepsilon \dfrac{d\psi(x)}{dx} \end{pmatrix} \tag{1.10}$$

and setting $\hbar = \varepsilon$, equation (1.2) becomes

$$i\varepsilon \frac{d\varphi(x)}{dx} = \begin{pmatrix} 0 & 1 \\ E - V(x) & 0 \end{pmatrix} \varphi(x). \tag{1.11}$$

Thus the three equations (1.7), (1.9), and (1.11) are particular cases of

$$i\varepsilon \frac{d\varphi(x)}{dx} = A(x)\varphi(x), \tag{1.12}$$

where $A(x)$ is a linear operator on $\mathcal{H} = \mathbf{C}^2$ with two distinct real eigenvalues. Our purpose is to study a scattering problem for (1.12) in the "semiclassical" limit ε tends to zero under the hypothesis that $A(x)$ is analytic, has two distinct real eigenvalues

for all $x \in \mathbf{R}$, and has well-defined limits when $x \to \pm\infty$. It is natural to express the solutions of (1.12) as linear combinations of eigenvectors of $A(x)$:

$$\varphi(x) = \sum_{i=1}^{2} c_j(x) e^{-i/\varepsilon \int_0^x e_j(x')\,dx'} \varphi_j(x), \tag{1.13}$$

where $A(x)\varphi_j(x) = e_j(x)\varphi_j(x)$. Our conditions on the behavior of $A(x)$ for large $|x|$ imply that

$$\lim_{x \to \pm\infty} c_j(x) = c_j(\pm\infty) \tag{1.14}$$

exist, so that the following scattering problem is well defined:

Given $c_j(-\infty), j = 1, 2$ find $c_j(+\infty), j = 1, 2$, i.e., find the matrix S defined by

$$\begin{pmatrix} c_1(+\infty) \\ c_2(+\infty) \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} c_1(-\infty) \\ c_2(+\infty) \end{pmatrix}. \tag{1.15}$$

There is a "canonical" choice of eigenvectors of $A(x)$ specified (up to a global factor) by the condition

$$P_j(x) \frac{d\varphi_j(x)}{dx} = 0, \tag{1.16}$$

where $P_j(x)$ is the eigenprojection corresponding to $e_j(x)$. Condition (1.16) has a geometrical interpretation in terms of parallel transport which we give below. In particular, it is immediate to verify that for $A(x)$ given by (1.9) or by (1.11) with the identification $\omega^2(x) \equiv E - V(x)$, the eigenvectors associated with $e_j(x) = (-1)^j \omega(x)$,

$$\varphi_1(x) = \begin{pmatrix} \dfrac{1}{\sqrt{\omega(x)}} \\ -\sqrt{\omega(x)} \end{pmatrix}, \qquad \varphi_2(x) = \begin{pmatrix} \dfrac{1}{\sqrt{\omega(x)}} \\ +\sqrt{\omega(x)} \end{pmatrix} \tag{1.17}$$

satisfy (1.16), so that (1.13) gives the solutions of (1.9) and (1.11) as superpositions of the two well-known Wentzel–Kramers–Brillouin (WKB) functions

$$e^{-i/\varepsilon \int_0^x e_j(x')\,dx'} \varphi_j(x). \tag{1.18}$$

When this choice of eigenvectors is made, a solution $\varphi(x)$ of (1.12) characterized by $c_j(-\infty) = 1$ and $c_k(-\infty) = 0, k \neq j$, satisfies

$$\sup_{x \in \mathbf{R}} |\varphi(x) - e^{-i/\varepsilon \int_0^x e_j(x')\,dx'} \varphi_j(x)| = O(\varepsilon). \tag{1.19}$$

Consequently,

$$S = \mathbf{1} + O(\varepsilon). \tag{1.20}$$

The approximations (1.19) and (1.20) are true without assuming analyticity of $A(x)$. On the other hand, if analyticity holds, we can approximate the solutions of (1.12) and

thus determine the matrix S up to error terms $O(\exp(-\kappa\varepsilon^{-1})), \kappa > 0$ (see Corollary 2.5),

$$S_{kj} = s_j(\varepsilon)\delta_{kj} + O(\exp(-\kappa\varepsilon^{-1})), \tag{1.21}$$

where $|s_j(\varepsilon)| = O(1)$. These results are corollaries of the iterative scheme presented in §2, which will be used in §3. Actually they are derived for $A(x)$ a $n \times n$ matrix whose eigenvalues are assumed to be real and nondegenerate for any $x \in \mathbf{R}$.

The asymptotic formulae (1.21) imply in particular that the nondiagonal terms of S are $O(\exp(-\kappa\varepsilon^{-1}))$. These terms are important in applications because they are related, for equation (1.1), to the probability of a quantum transition between the two levels of the system or, in the case of equation (1.2), to the above barrier reflection coefficient and, in the case of equation (1.4) to the quantity $\Delta J(\varepsilon)$. Under further hypotheses on the analytic behavior of the eigenvalues of $A(x)$ we show that it is possible to find an asymptotic expression for S_{21} or S_{12} accurate up to exponentially small *relative* corrections. The asymptotic formula is expressed by means of the complex degeneracy points of the analytic continuations of eigenvalues $e_j(x)$. If there are p contributing degeneracy points, the asymptotic expression reads (see Theorem 3.7 and (2.43), (2.45))

$$S_{21} = \sum_{k=0}^{p} e^{-i\theta^\star(k,\varepsilon)} e^{-i\gamma^\star(k,\varepsilon)\varepsilon^{-1}} + e^{-\tau\varepsilon^{-1}} O(e^{-\kappa\varepsilon^{-1}}), \qquad \kappa, \tau > 0, \tag{1.22}$$

where $\theta^\star(k,\varepsilon)$ is $O(1)$ and $\operatorname{Im}\gamma^\star(k,\varepsilon) = -\tau + O(\varepsilon^2), k = 1, \ldots, p$. It should be noted that the error term is smaller by an exponentially decreasing factor than the least significant term in the sum (1.22). This asymptotic formula is proven in §3, which is the main part of the paper. It is obtained by combining our iterative scheme with a method due to Fröman and Fröman [1]. We give in §4 explicit formulae in terms of $A(x)$ for the expressions $\theta^\star(k,\varepsilon)$ and $\gamma^\star(k,\varepsilon)$ appearing in (1.22). The consequences of our asymptotic analysis of the matrix S for the applications mentioned above are formulated in §4 as well. Finally, we give in the appendix an explicit example which is shown numerically to fit in the framework developed in this paper.

Let us come back to the choice of eigenvectors satisfying (1.16). Let M be some manifold, which we suppose to be embedded in $\mathbf{R}^q$, and let P be a smooth projection-valued map, $m \mapsto P(m)$, defined on $M, P(m)$ being a projection (not necessarily orthogonal) of some given Hilbert space. The map P defines a bundle F with base M, whose fiber over $m' \in M$ is the set of elements (m', ϕ) with $\phi \in P(m')\mathcal{H}$. The bundle F is embedded in the trivial bundle $\mathbf{R}^q \times \mathcal{H}$ and has a natural connection defined by P. Indeed, let $f = (m, \phi) \in F$; any tangent vector v_f at f can be viewed as a velocity vector of a curve $c(t) = (c_1(t), c_2(t))$ with $c_2(t) = P(c_1(t))c_2(t)$ and $c(0) = f$, i.e., $v_f = (\dot{c}_1(0), \dot{c}_2(0))_f$. The vertical vectors at f are velocity vectors of curves $c(t)$ with $c_1(t) \equiv m$; since in this case $c_2(t) \in P(m)\mathcal{H}$ for all t, they are of the form $(0, \dot{c}_2(0))_f$ with $\dot{c}_2(0) \in P(m)\mathcal{H}$. Conversely, since $c_2(t) = P(c_1(t))c_2(t)$, any vector of the form $(0, \dot{c}_2(0))_f$ is vertical. Therefore, we have a decomposition of v_f into a *vertical vector* $(0, P(m)\dot{c}_2(0))_f$ and a *horizontal vector* $(\dot{c}_1(0), (\mathbf{1} - P(m))\dot{c}_2(0))_f$, hence a *connection*. Let $t \mapsto \gamma(t)$ be a path in M and $\phi(t) \in P(\gamma(t))\mathcal{H}$ be a vector field along γ. This vector field is *parallel* if and only if the velocity vector $(\dot{\gamma}(t), \dot{\phi}(t))_{\gamma(t)}$ is horizontal for all t, i.e., if and only if $P(\gamma(t))\dot{\phi}(t) = 0$, which is precisely (1.16).

Before ending this introduction let us make some very brief comments on the vast amount of literature devoted to the exponential decay of nondiagonal elements of the

matrix S. We do not attempt at all to give an exhaustive account of it, but we want to set our work in context relative to the main results. We quote these results according to their content and not chronologically. The reader may find further references in the books [2] and [3]. The intermediate result (1.21) is not new, see [2], [3] and references therein, but we nevertheless obtain a new derivation of it in §2. For recent related results see also [4]. The asymptotic expression (1.22) generalizes several rigorous results which were obtained either in the case of equations (1.7) and (1.11) or in the study of $\Delta J(\varepsilon)$. When one complex eigenvalue degeneracy only contributes, it has been known since publication of the works [1], [5], [6] that

$$S_{21} = e^{-i\theta} e^{-i\gamma\varepsilon^{-1}} + O(\varepsilon) e^{\mathrm{Im}\gamma\varepsilon^{-1}}, \qquad \mathrm{Im}\gamma < 0 \tag{1.23}$$

with $\theta = \pi/2$ for equation (1.11) and, providing $A(x)$ is a real symmetric matrix, for equation (1.7) as well. It was shown recently that when $A(x)$ is a hermitian matrix in (1.7), θ can take any complex value [7], see also [8]. A corresponding asymptotic expression for $\Delta J(\varepsilon)$ in this situation can be found in [9]–[11]. See also [12] for more recent related results. The expression (1.23) was then generalized in two ways for equations (1.7) and (1.11). First, when several eigenvalue degeneracy points contribute to the asymptotics of S_{21}, it was proven using standard stretching and matching techniques that [5], [13]

$$S_{21} = \sum_{k=0}^{p} e^{-i\theta(k)} e^{-i\gamma(k)\varepsilon^{-1}} + O(\varepsilon^{\alpha}) e^{\mathrm{Im}\gamma\varepsilon^{-1}}, \tag{1.24}$$

where $0 < \alpha < 1$ and $\mathrm{Im}\gamma(k) = \mathrm{Im}\gamma < 0 \, \forall k$. The leading term of (1.24) gives rise to the so called "Stückelberg oscillations" as $\varepsilon \to 0$, a phenomenon which is illustrated numerically in [13]. Note also that the error term is $O(\varepsilon^{\alpha})$ instead of $O(\varepsilon)$, which is a common drawback of the method employed to get (1.24). Then, higher-order corrections to formula (1.23) were studied systematically in [14], [15] for equation (1.11) and in [16] for equation (1.7):

$$S_{21} = e^{-i\theta^q(\varepsilon)} e^{-i\gamma^q(\varepsilon)\varepsilon^{-1}} + O(\varepsilon^{q+1}) e^{-\tau\varepsilon^{-1}} \quad \forall q \in \mathrm{N}, \tau > 0, \tag{1.25}$$

where $\mathrm{Im}\gamma^q(\varepsilon) = -\tau + O(\varepsilon^2)$ and $\theta^q(\varepsilon) = O(1)$. The iterative scheme of §2 was introduced in [16] to derive this expression in the adiabatic context. Thus the asymptotic expression (1.22) captures all the features of these previous results and it holds for more general situations than those described by the particular matrices in (1.7) and (1.11). Moreover, it yields an expression accurate up to exponentially small corrections for the logarithm of S_{21} since we can write for $p = 1$

$$\ln S_{21} = -i\frac{\gamma^\star(\varepsilon)}{\varepsilon} - i\theta^\star(\varepsilon) + O(e^{-\kappa\varepsilon^{-1}}). \tag{1.26}$$

2. Approximate solution. The results of this section will be used in §3. We consider a slightly more general problem than in the introduction. Let $\mathcal{H} = \mathbf{C}^n$, with the usual scalar product, and $A(x), x \in \mathbf{R}$, be a linear operator on $\mathcal{H}$. We study the equation ($' = \frac{d}{dx}$)

$$\begin{aligned} i\varepsilon U'(x, x_0) &= A(x) U(x, x_0), \\ U(x_0, x_0) &= \mathbf{1}, \end{aligned} \tag{2.1}$$

under the condition that $A(x)$ is analytic in x and for each x the spectrum of $A(x)$ consists of n distinct real eigenvalues $e_1(x) < \cdots < e_n(x)$, with corresponding eigenprojections $P_1(x), \ldots, P_n(x)$. Note that the evolution U is not unitary in general.

In order to find an approximate solution of (2.1) we first consider another problem. Let $\psi(x)$ be a solution of

$$i\varepsilon\psi'(x) = A(x)\psi(x). \tag{2.2}$$

If $Q(x_0)$ is a projection such that $Q(x_0)\psi(x_0) = \psi(x_0)$, then for any x we have a projection $Q(x)$ such that $Q(x)\psi(x) = \psi(x)$. Indeed, if $U(x, x_0)$ is the solution of (2.1) such that $U(x_0, x_0) = \mathbf{1}$, we take

$$Q(x) = U(x, x_0)Q(x_0)U(x_0, x). \tag{2.3}$$

The projection $Q(x)$ is a solution of

$$i\varepsilon Q'(x) = [A(x), Q(x)] \tag{2.4}$$

with the notation $[A, B] \equiv AB - BA$. Let us suppose that at x_0 we have a complete set of projections $Q_j(x_0)$, i.e., $Q_j(x_0)Q_k(x_0) = Q_k(x_0)\delta_{jk}, \sum_j Q_j(x_0) = \mathbf{1}$. Then the $Q_j(x)$ form a complete set of projections as well and using the fact that for any projection $P(x)$ we have $P(x)P'(x)P(x) = 0$, it follows that

$$Q_j'(x) = \left[\sum_m Q_m'(x)Q_m(x), Q_j(x)\right]. \tag{2.5}$$

Therefore we have for all j

$$\left[A(x) - i\varepsilon \sum_m Q_m'(x)Q_m(x), Q_j(x)\right] = 0. \tag{2.6}$$

We look for approximate solutions of this equation. Since $[A(x), P_j(x)] \equiv 0$, the eigenprojections $P_j(x)$ are approximate solutions of (2.6) up to an error term $O(\varepsilon)$. Let

$$A_1(x) := A(x) - i\varepsilon K_0(x) \tag{2.7}$$

with

$$K_0(x) := \sum_m P_m'(x)P_m(x). \tag{2.8}$$

By perturbation theory, if ε is small enough, $A_1(x)$ has n distinct eigenvalues $e_{1,j}(x)$ with corresponding eigenprojections $P_{1,j}(x), j = 1, \ldots, n$, such that $e_{1,j}(x) = e_j(x) + O(\varepsilon^2)$, and $P_{1,j}(x) = P_j(x) + O(\varepsilon)$. Indeed, $e_{1,j}(x) = e_j(x) - i\varepsilon \operatorname{tr}(P_j(x)K_0(x)) + O(\varepsilon^2)$ and $P_j(x)K_0(x)P_j(x) = 0$. The projections $P_{1,j}(x)$ are approximate solutions of (2.6) up to an error term $O(\varepsilon^2)$ since $[A_1(x), P_{1,j}(x)] = 0$. Let

$$K_1(x) := \sum_m P_{1,m}'(x)P_{1,m}(x) \tag{2.9}$$

and

$$A_2(x) := A(x) - i\varepsilon K_1(x). \tag{2.10}$$

Again, for ε small enough, $A_2(x)$ has n distinct eigenvalues $e_{2,j}(x)$ with corresponding eigenprojections $P_{2,j}(x)$. Since $A_2(x) = A_1(x) + i\varepsilon(K_0(x) - K_1(x))$ and $K_0(x) - K_1(x) = O(\varepsilon), P_{2,j}(x)$ is an approximate solution of (2.6) up to an error term $O(\varepsilon^3)$. We can iterate this procedure. At the qth iteration we have approximate solutions $P_{q,j}(x)$, up to order term $O(\varepsilon^{q+1})$, which are eigenprojections of

$$A_q(x) := A(x) - i\varepsilon K_{q-1}(x) \tag{2.11}$$

with

$$K_{q-1}(x) = \sum_m P'_{q-1,m}(x) P_{q-1,m}(x). \tag{2.12}$$

We now construct approximate solutions for (2.1). Let $Q_m(x)$ be a complete smooth family of projections of $\mathcal{H}, Q_m(x)Q_n(x) = \delta_{mn}Q_m(x)$ and $\sum_m Q_m(x) = \mathbf{1}$. We say that an evolution $V(x,x'), (V(x',x') = \mathbf{1}, V(x_2,x_1)V(x_1,x_0) = V(x_2,x_0))$, follows the decomposition of $\mathcal{H}$,

$$\mathcal{H} = \bigoplus_m Q_m(x)\mathcal{H}$$

if for all x, x'

$$Q_m(x)V(x,x') = V(x,x')Q_m(x'). \tag{2.13}$$

It is known (see [17] or [18]) that a smooth evolution with property (2.13) is the solution of an equation of the type

$$V'(x,x_0) = \left(B(x) + \sum_m Q'_m(x)Q_m(x)\right) V(x,x_0), \qquad V(x_0,x_0) = \mathbf{1}, \tag{2.14}$$

where $B(x)$ is such that

$$[B(x), Q_m(x)] = 0 \quad \forall m. \tag{2.15}$$

Reciprocally, any smooth evolution satisfying (2.14) and (2.15) possesses the intertwining property (2.13). The idea is to construct approximate solutions of (2.1) by choosing evolutions which follow the decomposition of $\mathcal{H}$ into

$$\mathcal{H} = \bigoplus_m P_{q,m}(x)\mathcal{H}. \tag{2.16}$$

Therefore we define $U_q(x,x_0)$ as the solution of

$$i\varepsilon U'_q(x,x_0) = (A_q(x) + i\varepsilon K_q(x))U_q(x,x_0), \qquad U_q(x_0,x_0) = \mathbf{1} \tag{2.17}$$

The next lemma, which is actually Proposition 2.1 of [19], gives the main estimate which we need to control the error term for the approximate solution $U_q(x,x_0)$. This lemma is also used in §3.

For any $z \in \mathbf{C}$ and $r > 0$ let $D(z;r) = \{z' \in \mathbf{C} : |z'-z| < r\}$ and $\partial D(z;r) = \{z' \in \mathbf{C} : |z'-z| = r\}$. Given $z_0 \in \mathbf{C}$ and $r_0 > 0$ let $A(z)$ be analytic in $D(z_0;r_0)$ with a spectrum consisting of n distinct eigenvalues $e_j(z)$ with corresponding eigenprojection $P_j(z)$ for all $z \in D(z_0;r_0)$. We define $A_q(z), K_q(z), P_{q,j}(z)$, and $e_{q,j}(z)$ as above by the iteration method based on (2.11) and (2.12). We set $R(z,\lambda) = (A(z) - \lambda \mathbf{1})^{-1}$.

LEMMA 2.1. *Let $z_0 \in \mathbf{C}, r_0 > 0$ and $A(z)$ be defined on $D(z_0;r_0)$ with the above properties. Let $r_1 > 0$ and $D_j := D(e_j(z_0);2r_1)$ be n disjoint discs in $\mathbf{C}, j = 1,\dots,n$, such that for all $z \in D(z_0;r_0)$*

$$e_j(z) \in D(e_j(z_0);r_1).$$

Let

$$a = a(z_0) := \sup_j \sup_{\lambda \in \partial D_j} \sup_{z \in D(z_0;r_0)} \|R(z,\lambda)\| < \infty$$

and

$$b = b(z_0) := \sup_{z \in D(z_0;r_0)} \|K_0(z)\| < \infty.$$

Then there exist $\varepsilon^\star = \varepsilon^\star(a,b) > 0$ and $c = c(r_0,r_1,a,b) < \infty$ such that

$$\|K_q(z) - K_{q-1}(z)\| \le b\varepsilon^q c^q q!$$

and

$$\|K_q(z)\| \le 2b$$

for all $z \in D(z_0;r_0)$, all $0 < \varepsilon \le \varepsilon^\star$, and all $q \le q^\star(\varepsilon) = [\frac{1}{ec\varepsilon}]$, where $[y]$ is the integer part of y and e is the basis of the neperian logarithm.

Remark. The proof of this lemma is given in [19] for the case $P_1 + P_2 = \mathbf{1}$ in the general situation where the spectrum of the (possibly unbounded) operator $A(z)$ is separated in two parts for any $z \in D(z_0,r_0)$ and $\dim P_1(z)\mathcal{H} \le \infty$. However, the proof is the same for the case $\sum_{j=1}^n P_j = \mathbf{1}, n \ge 2$, apart from the obvious changes due to the presence of more than two projectors.

COROLLARY 2.2. *Let the hypothesis of Lemma 2.1 be satisfied. Then for all $q \le q^\star$*

$$e_{q,j}(z) = e_j(z) + O(b\varepsilon^2).$$

Proof. Since $P_j(z)K_0(z)P_j(z) = 0$ the statement is true for $q = 1$. For $q \ge 2$ we have

$$\|A_q(z) - A_1(z)\| \le \varepsilon \sum_{m=1}^{q-1} \|K_m(z) - K_{m-1}(z)\| \le \varepsilon b \sum_{m=1}^{q^*} \varepsilon^m c^m m! = O(\varepsilon^2 b) \tag{2.18}$$

and therefore the statement follows from perturbation theory. □

We now apply Lemma 2.1 and Corollary 2.2 to control the norm of $U_q(x,x_0)$. It is crucial that U_q follows the decomposition of $\mathcal{H}$ into $\bigoplus_{m\ge1} P_{q,m}(z)\mathcal{H}$.

COROLLARY 2.3. *Let $r_0 > 0$ be such that for each $x \in \mathbf{R}$ the hypotheses of Lemma 2.1 are satisfied on $D(x;r_0)$ with constants r_1 and a independent of x and with constants $b(x) \le b < \infty$. Then for $\varepsilon \le \varepsilon^\star$ and $q \le q^\star$*

$$\|U_q(x;x_0)\| \le \exp\left\{O\left(\left|\int_{x_0}^x b(x')\,dx'\right|\right)\right\}.$$

Proof. We introduce the evolution $W_q(x, x_0)$,

$$W_q'(x, x_0) = K_q(x) W_q(x, x_0), \qquad W_q(x_0, x_0) = \mathbf{1}. \tag{2.19}$$

From Lemma 2.1 we have

$$\|W_q(x, x_0)\| \leq \exp\left(2 \left| \int_{x_0}^{x} b(x')\, dx' \right|\right). \tag{2.20}$$

Let us choose n eigenvectors $\varphi_{q,j}(0)$ of $A_q(0)$ at $x = 0$. The vectors

$$\varphi_{q,j}(x) := W_q(x, 0)\varphi_{q,j}(0), \qquad j = 1, \ldots, n \tag{2.21}$$

are eigenvectors of $A_q(x)$ since $W_q(x, 0)$ interpolates between $P_{q,m}(0)$ and $P_{q,m}(x) \forall m \leq n$ (see (2.13) and (2.14)) and by definition

$$P_{q,j}(x)\varphi_{q,j}'(x) = 0, \qquad j = 1, \ldots, n. \tag{2.22}$$

Let us write $U_q(x, x_0) := W_q(x, x_0)\Phi_q(x, x_0)$. The unknown operator $\Phi_q(x, x_0)$ is the solution of

$$\begin{aligned} i\varepsilon \Phi_q'(x, x_0) &= W_q(x_0, x) A_q(x) W_q(x, x_0) \Phi_q(x, x_0), \\ \Phi_q(x_0, x_0) &= \mathbf{1}. \end{aligned} \tag{2.23}$$

The operator $W_q(x_0, x)A_q(x)W_q(x, x_0)$ has eigenvalues $e_{q,j}(x)$ with eigenvectors $\varphi_{q,j}(x_0)$. Therefore

$$\Phi_q(x, x_0)\varphi_{q,j}(x_0) = \exp\left(-i\varepsilon^{-1} \int_{x_0}^{x} e_{q,j}(x')\, dx'\right) \varphi_{q,j}(x_0), \qquad j = 1, \ldots, n. \tag{2.24}$$

From Corollary 2.2 and the reality of $e_j(z)$,

$$\left| \operatorname{Im}\left(\int_{x_0}^{x} e_{q,j}(x')\, dx' \right) \right| \leq O(\varepsilon^2) \left| \int_{x_0}^{x} b(x')\, dx' \right|, \tag{2.25}$$

hence

$$\|U_q(x, x_0)\| \leq \exp\left\{ (2 + O(\varepsilon)) \left| \int_{x_0}^{x} b(x')\, dx' \right| \right\}. \qquad \square \tag{2.26}$$

Note that in the above proof we have factorized the evolution $U_q(x, x_0)$ as the product

$$U_q(x, x_0) = W_q(x, x_0)\Phi_q(x, x_0), \tag{2.27}$$

where Φ_q only is singular in the limit $\varepsilon \to 0$ and $\|\Phi_q\| = O(1), \|W_q\| = O(1)$. Since in our simple case Φ_q is known explicitly, the solution $\psi(x)$ of

$$\begin{aligned} i\varepsilon \psi'(x) &= (A_q(x) + i\varepsilon K_q(x))\psi(x), \\ \psi(x_0) &= \psi_0 \end{aligned} \tag{2.28}$$

can be written as

$$\begin{aligned}\psi(x) &= U_q(x,x_0)\psi(0)\\ &= \sum_{j\geq 1} c_{q,j}(x_0)\exp\left(-i\,\varepsilon^{-1}\int_{x_0}^{x} e_{q,j}(x')\,dx'\right)\varphi_{q,j}(x),\end{aligned} \tag{2.29}$$

where the $c_{q,j}(x_0)$ are defined by the identity

$$\psi_0 = \sum_{j\geq 1} c_{q,j}(x_0)\varphi_{q,j}(x_0). \tag{2.30}$$

THEOREM 2.4. *Let $r > 0$ and $g > 0$ and let $A(x)$ be analytic in $\Omega_r = \{z = x+iy : x, y \in \mathbf{R}, |y| < r\}$. Let the spectrum of $A(x)$ consist of n real distinct eigenvalues $e_j(x), j = 1,\dots,n$, such that for all $x \in \mathbf{R}$*

$$|e_k(x) - e_j(x)| \geq g, \qquad k \neq j.$$

Let

$$\|K_0(x)\| = \left\|\sum_{j\geq 1} P_j'(x)P_j(x)\right\|$$

be an integrable function of x which tends to zero as $|x| \to 0$. Then there exist constants $\varepsilon^\star > 0, C' < \infty, \kappa > 0$ such that the above-constructed matrix $U_{q^\star}(x,x_0)$ approximates the solution $U(x,x_0)$ of the equation

$$\begin{aligned} i\varepsilon U'(x,x_0) &= A(x)U(x,x_0),\\ U(x_0,x_0) &= \mathbf{1}\end{aligned}$$

in such a way that

$$\sup_{x,x_0\in\mathbf{R}} \|U(x,x_0) - U_{q^\star}(x,x_0)\| \leq C'\exp(-\kappa\varepsilon^{-1}).$$

Remarks. i) Neither U nor $U_{q^\star}$ are unitary in general; however, both their norms are $O(1)$ as $\varepsilon \to 0$.

ii) Note that $\lim_{x\to\pm\infty} A(x)$ need not exist, since we only require that $\lim_{x\to\pm\infty} P_j(x) = P_j(\pm\infty)$ exists.

iii) The exponential decay rate is given by $\kappa = 1/ec$ (see (2.33)) where c is defined in Lemma 2.1. The decay rate obtained by this method is certainly not optimal but has the merit, however, to be explicit and rather simple to determine. It should be noted also that in the general case (i.e., $n > 2$), it is an open problem to determine the optimal decay rate.

iv) Similar results were also obtained by different methods: Nenciu [20] considered and studied a formal series expansion in ε satisfying (2.4) and Martinez [21] and Sjöstrand [22] used microlocal analysis techniques. In particular, the question raised in the preceding remark is addressed in [21]. However, the estimates needed in §3 are proved in [19] only.

Proof. By standard arguments of perturbation theory we can verify the hypothesis of Corollary 2.3 with $b(x)$ integrable on $\mathbf{R}$ (see, e.g., §2 of [23]). We recall that

$$q^\star(\varepsilon) = \left[\frac{1}{ec\varepsilon}\right] \tag{2.31}$$

as defined in Lemma 2.1. The operator $R(x) := U_{q^\star}(x_0, x)U(x, x_0)$ is a solution of

$$\begin{aligned} i\varepsilon R'(x) &= U_{q^\star}(x_0, x)(-A_{q^\star}(x) - i\varepsilon K_{q^\star}(x) + A(x))U_{q^\star}(x, x_0)R(x) \\ &= i\varepsilon U_{q^\star}(x_0, x)(K_{q^\star-1}(x) - K_{q^\star}(x))U_{q^\star}(x, x_0)R(x). \end{aligned} \tag{2.32}$$

From the integrability of $b(x)$ and Lemma 2.1 we have

$$\begin{aligned} \|R(x) - \mathbf{1}\| &\le C''(c\varepsilon)^{q^\star} q^\star! \\ &\le C''(c\varepsilon q^\star)^{q^\star} \\ &\le eC'' \exp(-\kappa\varepsilon^{-1}), \end{aligned} \tag{2.33}$$

where $\kappa = \frac{1}{\varepsilon c}$. Hence

$$\begin{aligned} \|U(x, x_0) - U_{q^\star}(x, x_0)\| &\le \|U_{q^\star}(x, x_0)\|\|R(x) - \mathbf{1}\| \\ &\le C' \exp(-\kappa\varepsilon^{-1}). \qquad \square \end{aligned} \tag{2.34}$$

We assume that the hypotheses of Theorem 2.4 are satisfied and we determine the matrix S up to an error term $O(e^{-\kappa\varepsilon^{-1}})$. Since $\|K_0(x)\|$ and thus $\|K_{q-1}(x)\|$ tend to zero at infinity in an integrable way (see Lemma 2.1 and Corollary 2.3),

$$\lim_{x\to\pm\infty} \|A_q(x) - A(x)\| = 0 \quad \forall q \le q^\star \tag{2.35}$$

and for all $q \le q^\star$, there exist $W_q(\pm\infty, x_0)$ such that

$$\lim_{x\to\pm\infty} W_q(x, x_0) = W_q(\pm\infty, x_0). \tag{2.36}$$

Let us choose a point x_0 and a set of eigenvectors $\varphi_j(x_0)$ of $A(x_0), j = 1, \dots, n$. Using $W_0(x, x_0)$ we define a set of eigenvectors of $A(x)$ for all x,

$$\varphi_j(x) = W_0(x, x_0)\varphi_j(x_0). \tag{2.37}$$

Let ψ be a solution of

$$i\varepsilon\psi'(x) = A(x)\psi(x) \tag{2.38}$$

and let us write ψ as

$$\psi(x) = \sum_{j\ge 1} c_j(x) e^{-i/\varepsilon \int_{x_0}^x e_j(x')\,dx'} \varphi_j(x). \tag{2.39}$$

Since $\|K_0(x)\|$ is integrable, $\lim_{x\to\pm\infty} c_j(x)$ exists (see, e.g., Lemma 3.2 below).

Let us now define a set of eigenvectors of $A_{q^\star}(x)$ by choosing

$$\varphi_j^\star(-\infty) \equiv \varphi_{q^\star,j}(-\infty) := \varphi_j(-\infty) \tag{2.40}$$

and setting

$$\varphi_j^\star(x) = W_{q^\star}(x, -\infty)\varphi_j(-\infty). \tag{2.41}$$

We can also write $\psi(x)$ as $(e_j^\star \equiv e_{q^\star,j})$

$$\begin{aligned}\psi(x) &= \sum_{j\geq 1} c_j^\star(x) e^{-i/\varepsilon \int_{x_0}^x e_j^\star(x')\,dx'} \varphi_j^\star(x) \\ &= \sum_{j\geq 1} c_j^\star(x) e^{-i/\varepsilon \int_{x_0}^x e_j(x')\,dx'} e^{-i/\varepsilon \int_{x_0}^x (e_j^\star(x')-e_j(x'))\,dx'} \varphi_j^\star(x). \end{aligned} \tag{2.42}$$

From (2.39), (2.42), and $\lim_{x\to-\infty} \|P_{q^\star,j}(x) - P_j(x)\| = 0$ we have

$$\begin{aligned}\lim_{x\to-\infty} e^{+i/\varepsilon \int_{x_0}^x e_j(x')\,dx'} P_j(x)\psi(x) &= c_j(-\infty)\varphi_j(-\infty) \\ &= e^{-i/\varepsilon \int_{x_0}^{-\infty} (e_j^\star(x')-e_j(x'))\,dx'} c_j^\star(-\infty)\varphi_j(-\infty). \end{aligned} \tag{2.43}$$

On the other hand, with the definitions $W_q(\pm\infty, \mp\infty) = W_q(\pm\infty, x_0)W_q(x_0, \mp\infty), 0 \leq q \leq q^\star$, we have

$$\begin{aligned}\varphi_j^\star(\infty) &= W_{q^\star}(\infty,-\infty)\varphi_j(-\infty) \\ &= W_{q^\star}(\infty,-\infty)W_0(-\infty,x_0)\varphi_j(x_0) \\ &= W_{q^\star}(\infty,-\infty)W_0(-\infty,\infty)\varphi_j(+\infty) \\ &\equiv e^{-i\beta_j^\star}\varphi_j(+\infty), \end{aligned} \tag{2.44}$$

the last equality defining the factor $e^{-i\beta_j^\star}$ where $\beta_j^\star$ is in general complex. Thus, similarly,

$$e^{-i/\varepsilon \int_{x_0}^{+\infty} (e_j^\star(x')-e_j(x'))\,dx'} e^{-i\beta_j^\star} c_j^\star(\infty) = c_j(\infty). \tag{2.45}$$

Let ψ be a solution of (2.38) characterized by $c_j(-\infty) = 1$ and $c_k(-\infty) = 0$ for $k \neq j$ which we decompose as in (2.42). From Theorem 2.4 and (2.29) an approximate solution of $\psi(x)$ is obtained by replacing $c_j^\star(x)$ by $c_j^\star(x_0)$ in (2.42), and we have

$$\sup_{x\in\mathbf{R}} |c_j^\star(x) - c_j^\star(x_0)| = O(e^{-\kappa\varepsilon^{-1}}), \qquad j = 1,\dots,n. \tag{2.46}$$

Therefore

$$c_k(+\infty) = O(e^{-\kappa\varepsilon^{-1}}), \qquad k \neq j \tag{2.47}$$

and

$$c_j(\infty) = e^{-i\beta_j^\star} e^{-i/\varepsilon \int_{-\infty}^{+\infty} (e_j^\star(x')-e_j(x'))\,dx'} + O(e^{-\kappa\varepsilon^{-1}}). \tag{2.48}$$

The matrix S defined in the introduction is then given by the following corollary.

COROLLARY 2.5.

$$S_{kj} = e^{-i\beta_j^\star} e^{-i/\varepsilon \int_{-\infty}^{+\infty} (e_j^\star(x')-e_j(x'))\,dx'} \delta_{kj} + O(e^{-\kappa\varepsilon^{-1}}).$$

Remark. It should be recalled that we did not write explicitly the ε-dependence of $e_j^\star$ or $P_{q^\star,j}$, but in Corollary 2.5 we have $\beta_j^\star = \beta_j^\star(\varepsilon)$ and $e_j^\star(x') = e_j^\star(x',\varepsilon)$.

3. Asymptotics of the nondiagonal part of the matrix S.

3.1. Stokes lines. From now on, we deal with the case $\mathcal{H} = \mathbf{C}^2$; we compute in this section an asymptotic expression for S_{21}, which, in the simplest case, reads

$$S_{21} = e^{-i\theta^\star(\varepsilon)} e^{-i\gamma^\star(\varepsilon)\varepsilon^{-1}} (1 + O(e^{-\kappa\varepsilon^{-1}})). \tag{3.1}$$

The idea is to combine our iterative scheme (2.11), (2.12) with an analysis in the complex plane by a method due to Fröman and Fröman [1]. To perform the analysis we need some precise information about the analytic extension of $A(x)$ into the complex plane. In particular, we must control the Stokes lines of the problem (Condition II below). Thus, in this subsection we introduce the notion of Stokes lines and give the conditions needed to make use of the method of [1] in the next subsection.

Without restricting the generality we impose $\mathrm{tr}A(x) \equiv 0$. Thus we have $A(x)^2 = \rho(x)\mathbf{1}$, with this identity defining the function $\rho(x)$. The eigenvalues of $A(x)$ are then $e_1(x) = -e_2(x)$ and $e_2(x) = \sqrt{\rho(x)}$, with $\sqrt{1} = 1$.

The corresponding eigenprojections are given by

$$P_j(x) = \frac{1}{2}\left(\mathbf{1} + \frac{A(x)}{e_j(x)}\right). \tag{3.2}$$

On $\mathbf{R}$ the eigenvalues are real and distinct and we suppose that there exists $g > 0$ with $\rho(x) > g$, for all $x \in \mathbf{R}$.

Let Ω be a domain of $\mathbf{C}$, symmetric with respect to the real axis, containing $\mathbf{R}$, on which A has an analytic extension. Since ρ is real on $\mathbf{R}$ we have for any $z \in \Omega, \rho(\bar{z}) = \overline{\rho(z)}$. The analysis of S_{21} is done by working in the upper half-plane only, whereas the analysis of S_{12} is performed in the lower half-plane, as we shall see below. The eigenvalues and eigenprojections also have analytic extensions in Ω, but it is clear that the zeros of ρ in Ω are singular points for these objects. Some of these singularities play a dominant role in the determination of $S_{jk}, j \neq k$.

As in §2 we introduce new operators $A_q(z)$ for all $z \in \Omega \setminus \{z' : \rho(z') = 0\}$ by the iteration scheme (2.11) and (2.12). In our case we can write

$$\begin{aligned} K_0(z) &= P_1'(z)P_1(z) + P_2'(z)P_2(z) \\ &= [P_1'(z), P_1(z)] = \frac{1}{4\rho(z)}[A'(z), A(z)], \end{aligned} \tag{3.3}$$

where $' = \frac{d}{dz}$ and we compute for all q

$$\begin{aligned} A_q(z) &= A(z) - i\varepsilon[P'_{q-1,1}(z), P_{q-1,1}(z)] \\ &= A(z) - \frac{i\varepsilon}{4\rho_{q-1}(z)}[A'_{q-1}(z), A_{q-1}(z)]. \end{aligned} \tag{3.4}$$

Indeed, we have $\mathrm{tr}A_{q'}(z) \equiv 0$, because the trace of a commutator is zero. Thus $\rho_{q'}(z)$ is defined by $A^2_{q'}(z) = \rho_{q'}(z)\mathbf{1}$. Hence the eigenvalues $e_{q',j}(z) = (-1)^j\sqrt{\rho_{q'}(z)}$ and $P_{q',1}(z)$ is given by an expression similar to (3.2). Equation (3.4) clearly shows that although the eigenvectors and eigenprojections are multivalued in Ω when we perform the analytic continuation, this is not the case for $A_q(z)$. In the above construction we must avoid the zeros of $\rho_{q'}(z)$ for $q' \leq q - 1$.

CONDITION I. *The set $X = \{z \in \Omega : \rho(z) = 0\}$ is a finite set. Let $r_2 > 0$ such that $D(z_j; r_2) \cap D(z_k; r_2) = \varnothing$ for all $z_j \neq z_k \in X$ and let*

$$\tilde{\Omega} = \Omega \setminus \bigcup_{z_j \in X} D(z_j; r_2). \tag{3.5}$$

There exist constants $g' > 0$ and $C' < \infty$ such that uniformly on $\tilde{\Omega}$

$$|\rho(z)| \geq g', \qquad \|P_j(z)\| \leq C'. \tag{3.6}$$

Remark. As we shall see in Conditions II and III below, we must satisfy (3.6) on a subset of $\tilde{\Omega}$ only.

Condition I allows us to verify the hypotheses of Lemma 2.1 uniformly on $\tilde{\Omega}$. Moreover the operators $A_q(z)$ are holomorphic on $\tilde{\Omega}$, provided ε is small enough. Indeed for any $\varepsilon \leq \varepsilon^\star$ and $q \leq q^\star$

$$\rho_q(z) = \rho(z) + O(b\varepsilon^2). \tag{3.7}$$

(The proof is the same as that of Corollary 2.2.) We define eigenvectors of $A_{q^\star}(z), z \in \tilde{\Omega}$, by the method of §2. Let $\varphi_j^\star(0)$ be an eigenvector of $A_{q^\star}(0)$ for the eigenvalue $e_j^\star(0), j = 1, 2$. Let $W_\star(z|\alpha)$ be the analytic continuation of $W_\star(x, 0)$ along a path α in $\tilde{\Omega}$, starting at 0 and ending at z, where

$$\begin{aligned} &W_\star'(x,0) = K_{q^\star}(x)W_\star(x,0), \qquad x \in \mathbf{R}, \\ &W_\star(0,0) = \mathbf{1}. \end{aligned} \tag{3.8}$$

The operator $W_\star(z|\alpha)$ is a (local) solution of

$$W_\star'(z|\alpha) = K_{q^\star}(z)W_\star(z|\alpha). \tag{3.9}$$

The main property of $W_\star(z|\alpha)$, which follows from (3.9) (see (2.13) and (2.14)), is that the vectors

$$\varphi_j^\star(z|\alpha) \equiv W_\star(z|\alpha)\varphi_j^\star(0), \qquad j = 1, 2 \tag{3.10}$$

are two eigenvectors of $A_{q^\star}(z)$, which are obtained by analytical continuation of $\varphi_j^\star(0)$ along α. The vector $\varphi_j^\star(z|\alpha)$ is an eigenvector for the eigenvalue $e_j^\star(z|\alpha)$, which is the analytic continuation of $e_j^\star(0)$ along α.

LEMMA 3.1. *Let z_j be a simple zero of ρ in Ω and let η be a simple closed path around $D(z_j; r_2)$, counterclockwise oriented and encircling no other disc $D(z_k; r_2)$ with $\rho(z_k) = 0$. Then for ε small enough,*

1) *the total variation of the argument of $\rho_{q^\star}$ along η is 2π, and*

2) *if η starts at $z = 0$, then there exist two complex numbers $\theta_{jk}^\star, j \neq k \quad j, k = 1, 2$, such that*

$$W_\star(0|\eta)\varphi_k^\star(0) := e^{i\theta_{jk}^\star}\varphi_j^\star(0), \qquad j \neq k$$

and

$$e^{i\theta_{kj}^\star}e^{i\theta_{jk}^\star} = -1, \qquad j \neq k.$$

Proof. 1) Using (3.7), we can write

$$\rho_{q^\star}(z) = \rho(z)g(z) \tag{3.11}$$

with $|g(z) - 1| < 1$ for all $z \in \eta$. Thus

$$\begin{aligned} 0 = \frac{1}{2\pi i}\int_\eta \frac{g'(z)}{g(z)}dz &= \frac{1}{2\pi i}\int_\eta \frac{\rho_{q^\star}'(z)}{\rho_{q^\star}(z)}dz - \frac{1}{2\pi i}\int_\eta \frac{\rho'(z)}{\rho(z)}dz \\ &= \frac{1}{2\pi i}\int_\eta \frac{\rho_{q^\star}'(z)}{\rho_{q^\star}(z)}dz - 1. \end{aligned} \tag{3.12}$$

2) $\varphi_j^\star(0)$ is an eigenvector of $A_{q^\star}(0)$ for the eigenvalue $e_j^\star(0)$. After analytical continuation $e_j^\star(0|\eta)$ is an eigenvalue of $A_{q^\star}(0)$ and by 1) it is equal to $-e_j^\star(0) = e_k^\star(0), k \neq j$. Thus $\varphi_j^\star(0|\eta) \equiv W_\star(0|\eta)\varphi_j^\star(0)$ is an eigenvector for the eigenvalue $e_k^\star(0)$ and therefore proportional to $\varphi_k^\star(0)$. Finally, the last identity is a consequence of $\det W_\star(z|\alpha) = 1$ since $\mathrm{tr} K_{q^\star}(z) \equiv 0$. □

Let Σ be a simply connected domain in $\tilde{\Omega}$, which contains the real axis. In Σ the analytic continuations of $e_j^\star(x)$ and $\varphi_j^\star(x)$ are path independent so that we write $e_j^\star(z)$ instead of $e_j^\star(z|\alpha)$ and so on. Let $\psi(z)$ be a solution of

$$i\varepsilon\psi'(z) = A(z)\psi(z), \qquad z \in \Sigma. \tag{3.13}$$

We decompose $\psi(z)$ along the eigenvectors of $A_{q^\star}(z)$,

$$\psi(z) = \sum_{j=1}^{2} c_j^\star(z) e^{-i/\varepsilon \int_0^z e_j^\star(z')\,dz'} \varphi_j^\star(z), \tag{3.14}$$

and we derive a differential equation for the unknown coefficients $c_j^\star(z)$ using the identities

$$A(z) = A_{q^\star}(z) + i\varepsilon K_{q^\star - 1}(z) \tag{3.15}$$

and

$$\varphi_j^{\star\prime}(z) = K_{q^\star}(z)\varphi_j^\star(z). \tag{3.16}$$

By performing scalar products with $W_\star^{-1}(z)^\dagger \varphi_j^\star(0), j = 1, 2$, where $\dagger$ denotes the adjoint, we get a set of linear equations to be solved for $c_j^{\star\prime}(z)$. Let R be the constant matrix defined by

$$R = \begin{pmatrix} \langle \varphi_1^\star(0)|\varphi_1^\star(0)\rangle & \langle \varphi_1^\star(0)|\varphi_2^\star(0)\rangle \\ \langle \varphi_2^\star(0)|\varphi_1^\star(0)\rangle & \langle \varphi_2^\star(0)|\varphi_2^\star(0)\rangle \end{pmatrix}^{-1}, \tag{3.17}$$

the elements of which, denoted by r_{jk}, are $O(1)$. We obtain finally

$$c_j^{\star\prime}(z) = \sum_{k=1}^{2} \exp(i\varepsilon^{-1}\Delta_{jk}^\star(z)) a_{jk}(z) c_k^\star(z), \tag{3.18}$$

where

$$\Delta_{jk}^\star(z) = \int_0^z (e_j^\star(z') - e_k^\star(z'))\,dz' \tag{3.19}$$

and

$$a_{jk}(z) = -\sum_{l=1}^{2} r_{jl} \langle \varphi_l^\star(0)|W_\star^{-1}(z)(K_{q^\star}(z) - K_{q^\star - 1}(z))W_\star(z)\varphi_k^\star(0)\rangle. \tag{3.20}$$

We have a good control of $a_{jk}(z)$ using Lemma 2.1 but the factor $\exp(i\varepsilon^{-1}\Delta_{jk}^\star(z))$ may cause trouble when we consider the limit $\varepsilon \to 0$ because $\mathrm{Im}\Delta_{jk}^\star(z) \neq 0$. Since $e_j^\star(z) = e_j(z) + O(\varepsilon^2 b)$, we must actually control the factor $\exp(i\varepsilon^{-1}\Delta_{jk}(z))$, where

$$\Delta_{jk}(z) = \int_0^z (e_j(z') - e_k(z'))\,dz'. \tag{3.21}$$

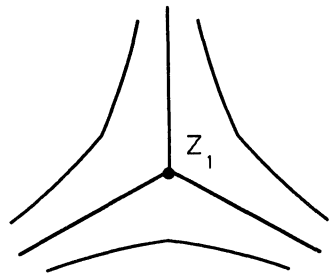


FIG. 1. *The level lines of* $\Phi(z)$ *near* z_0.

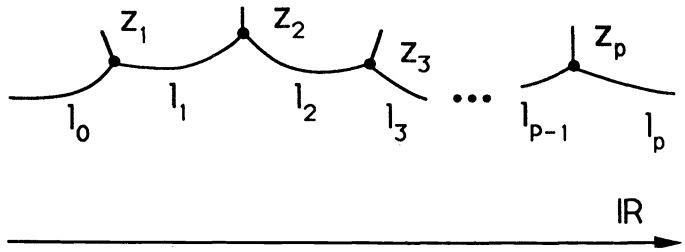


FIG. 2. *The Stokes lines of Condition* II.

The function Δ_{jk} is equal, up to a factor ± 2, to the function

$$\Phi(z) := \int_0^z \sqrt{\rho(z')}\, dz', \tag{3.22}$$

which is naturally associated with the quadratic differential $\rho(z)d^2z$.

DEFINITION. *A Stokes line* α *is a curve in* $\Omega \setminus \{z : \rho(z) = 0\}$ *such that*

1) $\mathrm{Im}\Phi(z)$ *is a constant along* α,

2) α *is maximal with property* 1), *and*

3) *one of the boundary points of* α *at least is a zero of* $\rho(z)$.

There are different terminologies in the literature. Sometimes our Stokes lines are called antiStokes lines and vice versa (see below). A Stokes line is always a simple curve and in our case it is contained either in the upper half-plane or in the lower half-plane. Near a simple zero z_0 of $\rho(z)$ the level-lines of $\mathrm{Im}\Phi(z)$ are homeomorphic to the level-lines

$$\mathrm{Im} z^{3/2} = \text{constant} \tag{3.23}$$

around $z = 0$. For any simple zero z_0 of $\rho(z)$ there are exactly three Stokes lines which have z_0 as boundary point. We call them the Stokes lines of z_0 (see Fig. 1).

CONDITION II. A) *There exists in the upper half-plane a nonempty finite set of simple zeros of* $\rho(z), \{z_1, \dots, z_p\}$ *with the following properties* (*see Fig.* 2):

1) *There exists a Stokes line* l_i, *parameterized by* (t_i, t_{i+1}), *such that* $\lim_{t\to t_i} l_i(t) = z_i, \lim_{t\to t_{i+1}} l_i(t) = z_{i+1}, i = 1, \dots, p-1$

2) *There exists a Stokes line* l_0, *parameterized by* $(-\infty, t_1)$, *such that* $\lim_{t\to t_1} l_0(t) = z_1, \lim_{t\to-\infty} \mathrm{Re} l_0(t) = -\infty, \lim_{t\to-\infty} \mathrm{Im} l_0(t) = a^-$

3) *there exists a Stokes line* l_p, *parameterized by* (t_p, ∞), *such that* $\lim_{t\to t_p} l_p(t) = z_p, \lim_{t\to\infty} \mathrm{Re} l_p(t) = \infty, \lim_{t\to\infty} \mathrm{Im} l_p(t) = a^+$.

B) *Along any vertical line* $\mathrm{Re} z = x$ *going from the real axis to* l_0 *or* $l_p, \mathrm{Im}\Phi(z)$ *is strictly monotone, provided* $|x|$ *is large enough.*

Remark. Condition II describes the situation illustrated in Fig. 2.

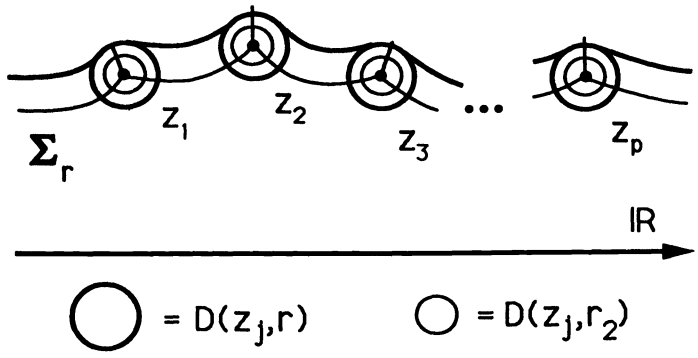


FIG. 3. *The set* Σ_r *of Condition* III.

In our case, if Condition II is satisfied then an analogous condition holds in the lower half-plane. It follows from Theorem 2.1 in [7] that the region Λ in the upper half-plane between the real axis and the closure of the Stokes lines $l_0, \dots, l_p$ is a simply connected region in Ω which does not contain zeros of ρ in its interior. In [7], part B of Condition II follows from the existence of limiting matrices when t tends to infinity. As already noted, such limiting matrices are not supposed to exist here. Let $r > 0$ and let

$$\Sigma_r = \{z \in \mathbf{C} \mid \operatorname{dist}(z, \Lambda) \leq r \text{ and } |z - z_i| \geq r, i = 1, \dots, p\}. \tag{3.24}$$

CONDITION III. *There exists* $r > r_2$, *sufficiently small so that* Σ_r *is a simply connected region in* Ω *containing the real axis and such that, for any zero* $z_i, i = 1, \dots, p$, *each Stokes line of* z_i *in the disc* $D(z_i; r)$ *intersects the boundary of the disc at a single point,* $D(z_i, r) \cap D(z_j, r) = \varnothing$ (*see Fig.* 3).

The function

$$b(x) := \sup_{\substack{y: \\ x+iy \in \Sigma_r}} \|K_0(x + iy)\| \tag{3.25}$$

tends to zero at infinity and is integrable on $\mathbf{R}$.

Remark. As we already mentioned, we need to verify Condition I on Σ_r only and not on $\tilde{\Omega}$ since we shall integrate the differential equation (3.18) along a path in Σ_r.

3.2. The Fröman–Fröman method. We suppose that Conditions I–III are satisfied and we study equation (3.18) on Σ_r. The hypotheses of Lemma 2.1 are thus verified uniformly on Σ_r, so that there exists a $q^\star = q^\star(\varepsilon)$ independent of $z \in \Sigma_r$ provided ε is small enough. Let us rewrite equation (3.18) as a Volterra equation

$$c_1^\star(z) = c_1^\star(z_0) + \int_{z_0}^{z} a_{11}(z') c_1^\star(z')\, dz' + \int_{z_0}^{z} a_{12}(z') e^{i\varepsilon^{-1}\Delta_{12}^\star(z')} c_2^\star(z')\, dz' \tag{3.26}$$

and

$$c_2^\star(z) = c_2^\star(z_0) + \int_{z_0}^{z} a_{22}(z') c_2^\star(z')\, dz' + \int_{z_0}^{z} a_{21}(z') e^{i\varepsilon^{-1}\Delta_{21}^\star(z')} c_1^\star(z')\, dz'. \tag{3.27}$$

LEMMA 3.2. *If Conditions* I–III *hold then* $\lim_{x \to \pm\infty} c_j^\star(x) = c_j^\star(\pm\infty)$ *exist and*

$$\lim_{x \to \pm\infty} \sup_{\substack{y: \\ x+iy \in \Sigma_r}} |c_j^\star(x + iy) - c_j^\star(\pm\infty)| = 0.$$

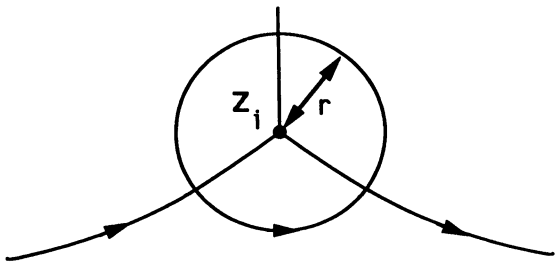


FIG. 4. *The path of integration close to* z_i.

Proof. By Conditions I–III we get from (3.20) and Lemma 2.1, as in §2,

$$\sup_{\substack{y:\\ x+iy\in\Sigma_r}} |a_{kj}(x+iy)| = b(x)O(e^{-\kappa\varepsilon^{-1}}) \tag{3.28}$$

and for all $z \in \Sigma_r$,

$$\Delta^\star_{jk}(z) = \Delta_{jk}(z) + O(\varepsilon^2). \tag{3.29}$$

Hence the limits $\lim_{x\to\pm\infty} c^\star_j(x)$ exist on the real axis since Δ_{jk} is real there. Then for all $z = x + iy$ on a vertical segment joining $\mathbf{R}$ and l_0 or l_p we can control $|\mathrm{Im}\Delta_{jk}(z)|$, provided $|x|$ is large enough, using part B of Condition II. Indeed, for such $z, |\mathrm{Im}\Delta_{jk}(z)|$ is bounded by twice the value of $|\mathrm{Im}\Phi(z)|$ on the Stokes lines. From these estimates and (3.28) we can easily deduce Lemma 3.2 using (3.26) and (3.27). □

Instead of integrating (3.18) along the real axis we integrate the equation along the Stokes lines $l_0, \dots, l_p$, as long as we are at a distance larger than r from a zero of ρ. Otherwise we integrate the equation along the boundaries of the discs $D(z_i; r)$, staying always in Σ_r (see Fig. 4).

Let z and z_0 be two points of Σ_r and let $T(z, z_0)$ be the matrix-solution of (3.18) with $T(z_0, z_0) = \mathbf{1}$. We can find $T(z, z_0)$ by integrating the equation along any path in Σ_r going from z_0 to z. However, because of the factors $\exp(i\varepsilon^{-1}\Delta_{jk}(z))$ we have a good control of the equation only on particular paths. For instance, the Stokes lines are "good" paths. The main work consists of controlling the equation along the parts of the boundaries of the discs $D(z_i; r)$ when we pass from one Stokes line to the next.

LEMMA 3.3. *Let z and $z_0 \in \Sigma_r$ and let α be a path, parameterized by $[s_0, s_1]$, going from z_0 to z, and such that $s \mapsto \mathrm{Im}\Delta_{12}(\alpha(s))$ is nondecreasing on $[s_0, s_1]$. Then*

$$T(z, z_0) = \begin{pmatrix} 1 + O(e^{-\kappa\varepsilon^{-1}}) & e^{-\varepsilon^{-1}\mathrm{Im}\Delta^\star_{12}(z_0)}O(e^{-\kappa\varepsilon^{-1}}) \\ e^{\varepsilon^{-1}\mathrm{Im}\Delta^\star_{12}(z)}O(e^{-\kappa\varepsilon^{-1}}) & 1 + O(e^{-\kappa\varepsilon^{-1}}) \\ & \quad + O(e^{-2\kappa\varepsilon^{-1}})e^{\varepsilon^{-1}(\mathrm{Im}\Delta^\star_{12}(z) - \mathrm{Im}\Delta^\star_{12}(z_0))} \end{pmatrix}.$$

Proof. We consider (3.26) and (3.27) along α with $c^\star_1(z_0) = 1$ and $c^\star_2(z_0) = 0$ and we introduce new variables

$$X_1(s) = c^\star_1(\alpha(s)), \qquad X_2(s) = e^{i\varepsilon^{-1}\Delta^\star_{12}(\alpha(s))}c^\star_2(\alpha(s)). \tag{3.30}$$

Writing $b_{jk}(s) = a_{jk}(\alpha(s))\frac{d\alpha(s)}{ds}$ and $\Delta^\star_{12}(s) \equiv \Delta^\star_{12}(\alpha(s))$, we get

$$X_1(s) = 1 + \int_{s_0}^{s} b_{11}(s')X_1(s')\,ds' + \int_{s_0}^{s} b_{12}(s')X_2(s')\,ds', \tag{3.31}$$

$$X_2(s) = \int_{s_0}^{s} b_{22}(s')e^{i\varepsilon^{-1}(\Delta^\star_{12}(s) - \Delta^\star_{12}(s'))}X_2(s')\,ds'$$

$$+ \int_{s_0}^{s} b_{21}(s')e^{i\varepsilon^{-1}(\Delta^\star_{12}(s) + \Delta^\star_{21}(s'))}X_1(s')\,ds'. \tag{3.32}$$

In (3.32) $s' \leq s$ and $\Delta_{21}^\star(s') = -\Delta_{12}^\star(s')$. Using (3.29) and the hypothesis on the path we have

$$\begin{aligned} &|e^{i\varepsilon^{-1}(\Delta_{12}^\star(s)-\Delta_{12}^\star(s'))}| \\ &\quad = \exp(-\varepsilon^{-1}(\mathrm{Im}\Delta_{12}(s) - \mathrm{Im}(\Delta_{12}(s'))) + O(\varepsilon)) = O(\exp(O(\varepsilon))). \end{aligned} \tag{3.33}$$

Let $\|X_i\| = \sup_{s_0 \leq s \leq s_1} |X_i(s)|$. We get from (3.31), (3.32), and (3.33), using (3.28),

$$\begin{aligned} \|X_1\| &\leq 1 + O(e^{-\kappa\varepsilon^{-1}})(\|X_1\| + \|X_2\|), \\ \|X_2\| &\leq O(e^{-\kappa\varepsilon^{-1}})(\|X_1\| + \|X_2\|), \end{aligned} \tag{3.34}$$

so that for ε small enough $\|X_1\| + \|X_2\| \leq 2$. Using this a priori estimate in (3.31) and (3.32) we have

$$\sup_{s_0 \leq s \leq s_1} |X_1(s) - 1| = O(e^{-\kappa\varepsilon^{-1}}) \tag{3.35}$$

and

$$\sup_{s_0 \leq s \leq s_1} |X_2(s)| = O(e^{-\kappa\varepsilon^{-1}}). \tag{3.36}$$

Equations (3.35) and (3.36) allow us to determine the first column of $T(z, z_0)$,

$$T(z, z_0) = \begin{pmatrix} 1 + O(e^{-\kappa\varepsilon^{-1}}) & T_{12}(z, z_0) \\ e^{\varepsilon^{-1}\mathrm{Im}\Delta_{12}^\star(z)} O(e^{-\kappa\varepsilon^{-1}}) & T_{22}(z, z_0) \end{pmatrix}. \tag{3.37}$$

Since $|a_{11}(z) + a_{22}(z)| = O(e^{-\kappa\varepsilon^{-1}})$, we get from the Liouville formula

$$\begin{aligned} \det T(z, z_0) &= \exp(O(e^{-\kappa\varepsilon^{-1}})) \\ &= 1 + O(e^{-\kappa\varepsilon^{-1}}). \end{aligned} \tag{3.38}$$

Moreover $T^{-1}(z, z_0) = T(z_0, z)$, hence

$$T(z_0, z) = \frac{1}{\det T(z, z_0)} \begin{pmatrix} T_{22}(z, z_0) & -T_{12}(z, z_0) \\ -T_{21}(z, z_0) & T_{11}(z, z_0) \end{pmatrix}. \tag{3.39}$$

The reverse path α^{-1} from z to z_0 is such that $s \mapsto \mathrm{Im}\Delta_{21}(\alpha^{-1}(s))$ is nonincreasing from s_1 to s_0. If $c_1^\star(z) = 0$ and $c_2^\star(z) = 1$ then we can estimate $c_1^\star(z_0)$ and $c_2^\star(z_0)$ as above, introducing new variables $Y_2(s) = c_2^\star(\alpha^{-1}(s))$ and $Y_1(s) = e^{i\varepsilon^{-1}\Delta_{21}^\star(\alpha^{-1}(s))} \times c_1^\star(\alpha^{-1}(s))$. Thus we can estimate the second column of (3.39). The coefficient $T_{22}(z, z_0)$ is estimated using $\det T(z, z_0) = 1 + O(e^{-\kappa\varepsilon^{-1}})$. □

A Stokes line is a good path because $\mathrm{Im}\Delta_{jk}(z)$ remains constant along this line. The following corollary is thus immediate.

COROLLARY 3.4. *If there is a Stokes line going from z_0 to z, then*

$$T(z, z_0) = \begin{pmatrix} 1 + O(e^{-\kappa\varepsilon^{-1}}) & O(e^{-\kappa\varepsilon^{-1}}) e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}^\star(z)} \\ O(e^{-\kappa\varepsilon^{-1}}) e^{\varepsilon^{-1}\mathrm{Im}\Delta_{12}^\star(z)} & 1 + O(e^{-\kappa\varepsilon^{-1}}) \end{pmatrix}.$$

We now come to the difficult part of the method. We must control the matrix solution $T(z, z_0)$ along a portion of $\partial D(z_j, r)$, which is not a good path in the sense that $\mathrm{Im}\Delta_{12}(z)$ is not monotone. We must establish two lemmas. The first lemma

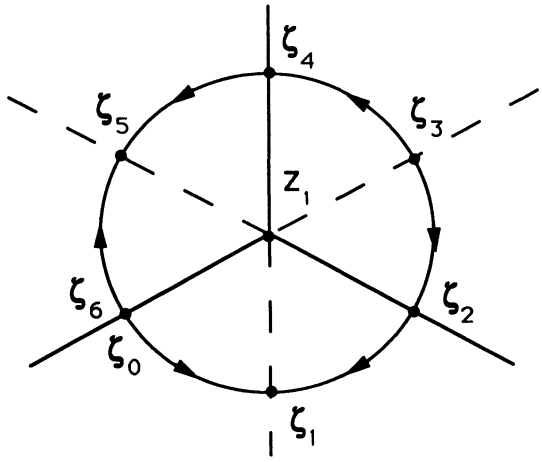


FIG. 5. *The points $\zeta_j, j = 0, \ldots, 6$ on the Stokes and antiStokes lines.*

gives a monodromy matrix around the singularity z_1 and is easily proven. The second and main lemma is more difficult to establish. Its proof is based on Lemmas 3.3 and 3.5 and on a clever use of elementary identities between the coefficients of products of 2×2 matrices and their inverses [1]. This method has a definite advantage over the use of stretching and matching techniques to compute asymptotics in the sense that it allows us to obtain better estimates on the remainders (see (1.19) in the introduction). However, it can only be used for simple zeros of the function $\rho(z)$, whereas the stretching and matching method works in more general situations [24].

We consider now the neighborhood of a zero of $\rho(z)$, say z_1. Let δ be the boundary of the disc $D(z_1; r)$ counterclockwise oriented, going from ζ_0 to ζ_6 as in Fig. 5. On this figure the solid lines are the Stokes lines of z_1 and the dashed lines are the antiStokes lines of z_1, i.e., the lines along which $\mathrm{Re}\Delta_{12}(z) \equiv \mathrm{Re}\Delta_{12}(z_1)$. The arrows indicate the directions in which $\mathrm{Im}\Delta_{12}(z)$ is nondecreasing along the boundary of $D(z_1; r)$.

We compute the matrix $T(\zeta_6, \zeta_0)$ along δ.

LEMMA 3.5.

$$T(\zeta_6, \zeta_0) = \begin{pmatrix} 0 & e^{i/\varepsilon \int_\eta e_1^\star} e^{-i\theta_{21}^\star} \\ e^{i/\varepsilon \int_\eta e_2^\star} e^{-i\theta_{12}^\star} & 0 \end{pmatrix}.$$

Proof. Let us consider $\psi(z)$ at $z = \zeta_0$, the solution of which we have obtained by integration along the Stokes line l_0 up to ζ_0. We have

$$\psi(\zeta_0) = \sum_{j=1}^{2} c_j^\star(\zeta_0) e^{-i/\varepsilon \int_0^{\zeta_0} e_j^\star} \varphi_j^\star(\zeta_0), \tag{3.40}$$

where in (3.40) the integration from 0 to ζ_0 is along α as in Fig. 6 and, similarly, $\varphi_j^\star(\zeta_0)$ is the analytical continuation of $\varphi_j^\star(0)$ along α.

We make the analytical continuation of (3.40) along δ up to ζ_6. Since $\psi(z)$ is holomorphic at z_1 we have $\psi(\zeta_6) = \psi(\zeta_0)$ and we can write

$$\psi(\zeta_0) = \sum_{j=1}^{2} c_j^\star(\zeta_6) e^{-i/\varepsilon \int_\alpha e_j^\star} e^{-i/\varepsilon \int_\delta e_j^\star} \varphi_j^\star(\zeta_6), \tag{3.41}$$

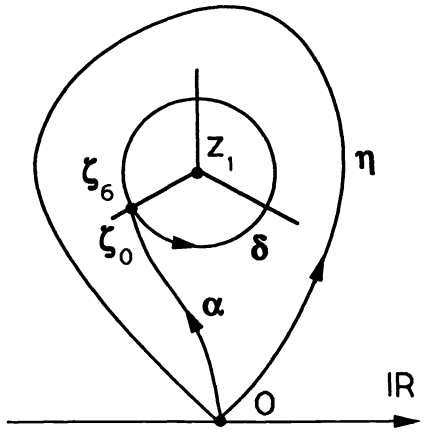


FIG. 6. *The paths α, δ, and η.*

where now $\varphi_j^\star(\zeta_6)$ is the analytical continuation of $\varphi_j^\star(0)$ along α and then along δ. But this is the same as the analytical continuation of $\varphi_j^\star(0)$ along η and then along α as in Fig. 6. By Lemma 3.1 we therefore have

$$\varphi_j^\star(\zeta_6) = e^{i\theta_{kj}^\star}\varphi_k^\star(\zeta_0). \tag{3.42}$$

Similarly we have

$$\int_\alpha e_j^\star + \int_\delta e_j^\star = \int_\eta e_j^\star + \int_\alpha e_k^\star. \tag{3.43}$$

Hence, by comparing (3.40) and (3.41),

$$c_j^\star(\zeta_6)e^{-i/\varepsilon\int_\eta e_j^\star}e^{i\theta_{kj}^\star} = c_k^\star(\zeta_0), \qquad k \neq j. \tag{3.44}$$ □

LEMMA 3.6. *For ε small enough*

$$T(\zeta_2,\zeta_0) = \begin{pmatrix} 1+O(e^{-\kappa/\varepsilon}) & O(e^{-\kappa/\varepsilon})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(\zeta_0)} \\ e^{-i/\varepsilon\int_\eta e_1^\star}e^{-i\theta_{12}^\star}(1+O(e^{-\kappa/\varepsilon})) & 1+O(e^{-\kappa/\varepsilon}) \end{pmatrix}.$$

Proof. The following computations will involve expressions such as $\varepsilon^{-1}\mathrm{Im}\Delta_{12}^\star(\zeta_\nu)$ for $\nu = 0,2,4,6$. These expressions are almost equal. Indeed

$$\Delta_{jk}^\star(z) = \Delta_{jk}(z) + O(\varepsilon^2) \tag{3.45}$$

and for this choice of ζ_ν we have

$$\mathrm{Im}\Delta_{12}(\zeta_\nu) = \mathrm{Im}\Delta_{12}(z_1), \qquad \nu = 0,2,4,6 \tag{3.46}$$

since these points are on the Stokes lines of z_1. Hence, in particular,

$$e^{\pm\varepsilon^{-1}\mathrm{Im}\Delta_{12}^\star(\zeta_\nu)} = O(e^{\pm\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}), \qquad \nu = 0,2,4,6. \tag{3.47}$$

Finally note that (see Fig. 6)

$$\int_\eta e_1^\star = \int_\eta e_1 + O(\varepsilon^2) = \Delta_{12}(z_1) + O(\varepsilon^2). \tag{3.48}$$

Let us denote the coefficient jk of the matrix $T(\zeta_\alpha, \zeta_\beta)$ by $t_{jk}(\alpha, \beta)$ and consider the identity

$$T(\zeta_{\nu+1}, \zeta_\nu) = T(\zeta_{\nu+1}, \zeta_{\nu+2})T(\zeta_{\nu+2}, \zeta_\nu). \tag{3.49}$$

Using (3.38) again

$$\det T(\zeta_\mu, \zeta_\nu) = t_{11}(\mu, \nu)t_{22}(\mu, \nu) - t_{12}(\mu, \nu)t_{21}(\mu, \nu) = 1 + O(e^{-\kappa\varepsilon^{-1}}) \tag{3.50}$$

and we obtain for $\nu = 0, 2$, and 4

$$t_{11}(\nu+2, \nu) = \frac{t_{11}(\nu+1, \nu)}{t_{11}(\nu+1, \nu+2)} - \frac{t_{12}(\nu+1, \nu+2)}{t_{11}(\nu+1, \nu+2)} t_{21}(\nu+2, \nu), \tag{3.51}$$

$$\begin{aligned} t_{22}(\nu+2, \nu) &= \frac{t_{11}(\nu+1, \nu+2)}{t_{11}(\nu+1, \nu)}(1 + O(e^{-\kappa\varepsilon^{-1}})) \\ &\quad + \frac{t_{12}(\nu+1, \nu)}{t_{11}(\nu+1, \nu)} t_{21}(\nu+2, \nu), \end{aligned} \tag{3.52}$$

$$\begin{aligned} t_{12}(\nu+2, \nu) &= \frac{t_{12}(\nu+1, \nu)}{t_{11}(\nu+1, \nu+2)} - \frac{t_{12}(\nu+1, \nu+2)}{t_{11}(\nu+1, \nu)}(1 + O(e^{-\kappa\varepsilon^{-1}})) \\ &\quad - \frac{t_{12}(\nu+1, \nu)t_{12}(\nu+1, \nu+2)}{t_{11}(\nu+1, \nu)t_{11}(\nu+1, \nu+2)} t_{21}(\nu+2, \nu). \end{aligned} \tag{3.53}$$

These identities express, in particular, the elements of the matrix $T(\zeta_2, \zeta_0)$ as functions of the element $t_{21}(2, 0)$ and other matrix elements that we can control by means of Lemma 3.3:

$$t_{11}(2, 0) = 1 + O(e^{-\kappa\varepsilon^{-1}}) + O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}t_{21}(2, 0), \tag{3.54}$$

$$t_{22}(2, 0) = 1 + O(e^{-\kappa\varepsilon^{-1}}) + O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}t_{21}(2, 0), \tag{3.55}$$

$$t_{12}(2, 0) = O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)} + (O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)})^2 t_{21}(2, 0). \tag{3.56}$$

We are thus led to the determination of $t_{21}(2, 0)$. Note that these estimates are true for the elements of $T(\zeta_6, \zeta_4)$ if we replace the arguments $(2, 0)$ by $(6, 4)$. Consider now the identity

$$T(\zeta_3, \zeta_2)T(\zeta_2, \zeta_0)T(\zeta_0, \zeta_6) = T(\zeta_3, \zeta_4)T(\zeta_4, \zeta_6). \tag{3.57}$$

Using Lemma 3.1 and $e_1^\star \equiv -e_2^\star$ to compute $T(\zeta_0, \zeta_6) = T(\zeta_6, \zeta_0)^{-1}$, we obtain for the coefficient 22 of (3.57)

$$\begin{aligned} &t_{21}(3, 2)t_{11}(2, 0)e^{i\theta_{12}^\star}e^{i\varepsilon^{-1}\int_\eta e_1^\star} + t_{22}(3, 2)t_{21}(2, 0)e^{i\theta_{12}^\star}e^{i\varepsilon^{-1}\int_\eta e_1^\star} \\ &\quad = t_{21}(3, 4)t_{12}(4, 6) + t_{22}(3, 4)t_{22}(4, 6) \end{aligned} \tag{3.58}$$

and for the coefficient 21 of (3.57)

$$\begin{aligned} &t_{21}(3, 2)t_{12}(2, 0)e^{i\theta_{21}^\star}e^{i\varepsilon^{-1}\int_\eta e_2^\star} + t_{22}(3, 2)t_{22}(2, 0)e^{i\theta_{21}^\star}e^{i\varepsilon^{-1}\int_\eta e_2^\star} \\ &\quad = t_{21}(3, 4)t_{11}(4, 6) + t_{22}(3, 4)t_{21}(4, 6). \end{aligned} \tag{3.59}$$

Lemma 3.3, (3.39), and (3.47) yield

$$t_{21}(3,2) = -t_{21}(2,3)(1+O(e^{-\kappa\varepsilon^{-1}})) = O(e^{-\kappa\varepsilon^{-1}})e^{\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}, \tag{3.60}$$

$$t_{21}(3,4) = -t_{21}(4,3)(1+O(e^{-\kappa\varepsilon^{-1}})) = O(e^{-\kappa\varepsilon^{-1}})e^{\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}, \tag{3.61}$$

$$t_{22}(3,4) = t_{11}(4,3)(1+O(e^{-\kappa\varepsilon^{-1}})) = 1+O(e^{-\kappa\varepsilon^{-1}}), \tag{3.62}$$

$$t_{22}(3,2) = t_{11}(2,3)(1+O(e^{-\kappa\varepsilon^{-1}})) = 1+O(e^{-\kappa\varepsilon^{-1}}), \tag{3.63}$$

whereas from (3.39) and the remark following (3.56) we have

$$t_{12}(4,6) = O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)} + (O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)})^2 t_{21}(6,4) \tag{3.64}$$

and

$$t_{22}(4,6) = 1+O(e^{-\kappa\varepsilon^{-1}}) + O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}t_{21}(6,4). \tag{3.65}$$

Now we use (3.58) and the above results to get

$$\begin{aligned} t_{21}(2,0)e^{i\theta_{12}^\star}e^{i\varepsilon^{-1}\int_\eta e_1^\star} &= 1+O(e^{-\kappa\varepsilon^{-1}}) + O(e^{-2\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}t_{21}(2,0) \\ &\quad + O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}t_{21}(6,4). \end{aligned} \tag{3.66}$$

Hence we see that we have to estimate $t_{21}(6,4)$ as well as determine $t_{21}(2,0)$. This is done by performing a similar computation: We estimate $t_{11}(4,6)$ as a function of $t_{21}(6,4)$ as above and we consider equation (3.59). After multiplication by $e^{-i\theta_{21}^\star} \times e^{-i\varepsilon^{-1}\int_\eta e_2^\star}$ and using

$$\mathrm{Im}\int_\eta e_2^\star = -\mathrm{Im}\int_\eta e_1^\star, \tag{3.67}$$

we obtain another equation for $t_{21}(6,4)$ and $t_{21}(2,0)$:

$$\begin{aligned} &-t_{21}(6,4)e^{-i\theta_{21}^\star}e^{-i\varepsilon^{-1}\int_\eta e_2^\star} = 1+O(e^{-\kappa\varepsilon^{-1}}) \\ &+O(e^{-2\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}t_{21}(6,4) + O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}t_{21}(2,0). \end{aligned} \tag{3.68}$$

Therefore, from (3.66) and (3.68) we deduce the a priori estimates

$$e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}|t_{21}(2,0)| = O(1), \tag{3.69}$$

$$e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)}|t_{21}(6,4)| = O(1), \tag{3.70}$$

which finally yield

$$t_{21}(2,0) = e^{-i\theta_{12}^\star}e^{-i\varepsilon^{-1}\int_\eta e_1^\star}(1+O(e^{-\kappa\varepsilon^{-1}})). \tag{3.71}$$ □

This lemma and Corollary 3.4 allow us to obtain an asymptotic expression for $\ln S_{21}$ beyond all orders by integrating (3.18) from $-\infty$ to $+\infty$ along the paths described above. Let us recall that we have

$$\mathrm{Im}\Delta_{12}(z_1) \equiv \mathrm{Im}\Delta_{12}(z_i), \qquad i=1,\dots,p. \tag{3.72}$$

Thus, along the Stokes lines we use the matrices given by Corollary 3.4 and which we can write as

$$T := T(z, z_0) = \begin{pmatrix} 1 + O(e^{-\kappa\varepsilon^{-1}}) & O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)} \\ O(e^{-\kappa\varepsilon^{-1}})e^{\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)} & 1 + O(e^{-\kappa\varepsilon^{-1}}) \end{pmatrix}. \tag{3.73}$$

On the other hand, when we go from one Stokes line, l_{j-1}, to the next one, l_j, we use the matrix given by Lemma 3.6:

$$S_j := \begin{pmatrix} 1 + O(e^{-\kappa\varepsilon^{-1}}) & O(e^{-\kappa\varepsilon^{-1}})e^{-\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)} \\ e^{-i/\varepsilon \int_{\eta_j} e_1^\star} e^{-i\theta_{12}^\star(j)}(1 + O(e^{-\kappa\varepsilon^{-1}})) & 1 + O(e^{-\kappa\varepsilon^{-1}}) \end{pmatrix}, \tag{3.74}$$

where $\int_{\eta_j} e_1^\star$ and $\theta_{21}^\star(j)$ are the quantities associated with the simple zero z_j of $\rho(z)$. Therefore if we start at $-\infty$ with the values $c_1^\star(-\infty) = 1$ and $c_2^\star(0) = 0$, then the coefficients $c_1^\star(+\infty)$ and $c_2^\star(+\infty)$ are obtained by computing

$$\begin{pmatrix} c_1^\star(\infty) \\ c_2^\star(\infty) \end{pmatrix} = TS_pTS_{p-1} \dots S_1T \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{3.75}$$

which proves the final theorem of this section, (restoring the ε dependence):

THEOREM 3.7. *Under Conditions* I–III, *the solution of* (3.18) *such that* $c_1^\star(-\infty) = 1$ *and* $c_2^\star(-\infty) = 0$ *is given at* $x = +\infty$ *by*

$$c_1^\star(\infty) = 1 + O(e^{-\kappa\varepsilon^{-1}})$$

and

$$c_2^\star(\infty) = \sum_{k=1}^{p} e^{-i/\varepsilon \int_{\eta_k} e_1^\star(z,\varepsilon)\, dz} e^{-i\theta_{12}^\star(k,\varepsilon)} + O(e^{-\kappa\varepsilon^{-1}})e^{\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)},$$

where $\mathrm{Im} \int_{\eta_k} e_1^\star(z,\varepsilon)\, dz = \mathrm{Im}\Delta_{12}(z_1) + O(\varepsilon^2)$ *and* $\theta_{12}^\star(k,\varepsilon) = O(1)$.

4. Applications.

4.1. Explicit formulae. Let us start by deriving explicit formulae for the eigenvectors $\varphi_j^\star(z)$ of $A_{q^\star}(z)$ defined by (3.10). They will then allow us to give the precise relation between the coefficients $c_j(z)$ defined by the expansion

$$\varphi(z) = \sum_{j=1}^{2} c_j(z) e^{-i/\varepsilon \int_0^z e_j(z')\, dz'} \varphi_j(z) \tag{4.1}$$

and the coefficients $c_j^\star(z)$ defined by

$$\varphi(z) = \sum_{j=1}^{2} c_j^\star(z) e^{-i/\varepsilon \int_0^z e_j^\star(z')\, dz'} \varphi_j^\star(z). \tag{4.2}$$

Note that here we have chosen $z_0 = 0$. Consider the operator $A_{q^\star}(z), z \in \Sigma$, where Σ is a simply connected domain of $\tilde{\Omega}$. We can write

$$A_{q^\star}(z) = \begin{pmatrix} ic^\star(z) & a^\star(z) \\ b^\star(z) & -ic^\star(z) \end{pmatrix} \tag{4.3}$$

with

$$\rho_{q^\star}(z) \equiv \rho_\star(z) = a_\star(z)b_\star(z) - (c_\star(z))^2. \tag{4.4}$$

LEMMA 4.1. *The eigenvectors of $A_{q^\star}(z)$ defined by* (3.10) *are given by*

$$\varphi_j^\star(z) = \frac{\chi_j^\star(z)}{\|\chi_j^\star(-\infty)\|} e^{-i(-1)^j \sigma_\star(z)}, \qquad j = 1, 2,$$

where

$$\chi_j^\star(z) = \begin{pmatrix} \sqrt{\frac{a_\star(z)}{\sqrt{\rho_\star(z)}}} \\ (-1)^j \sqrt{\frac{\sqrt{\rho_\star(z)}}{a_\star(z)}} - i \frac{c_\star(z)}{\sqrt{\sqrt{\rho_\star(z)} a_\star(z)}} \end{pmatrix}$$

and

$$\sigma_\star(z) = \frac{1}{2} \int_{-\infty}^{z} \frac{c_\star(u)a_\star'(u) - c_\star'(u)a_\star(u)}{\sqrt{\rho_\star(u)} a_\star(u)} du$$

for any $z \in \Sigma \setminus Y_\star$ and $Y_\star = \{z \in \Sigma : a_\star(z) = 0\}$.

Remarks. i) Any traceless matrix can be written under the form given above; the lemma actually requires the existence of distinct eigenvalues only. It is true in particular for the operator $A(z)$ written as in (4.3) without indices $\star$.

ii) The vectors $\varphi_j^\star(z)$ are actually analytic in the whole set Σ since the operator $W_\star(z)$ is analytic in Σ.

Proof. A direct verification shows that the vectors $\chi_j^\star(z)$ are eigenvectors of $A_{q^\star}(z)$ for the eigenvalues $e_j^\star(z) = (-1)^j \sqrt{\rho_\star(z)}$. We set the notation

$$p_\star(z) = \sqrt{\rho_\star(z)} \tag{4.5}$$

and we introduce the eigenprojectors (see (3.2))

$$P_{q^\star,j}(z) \equiv P_j^\star(z) = \frac{1}{2} \begin{pmatrix} 1 + (-1)^j \frac{ic_\star(z)}{p_\star(z)} & (-1)^j \frac{a_\star(z)}{p_\star(z)} \\ (-1)^j \frac{b_\star(z)}{p_\star(z)} & 1 - (-1)^j \frac{ic_\star(z)}{p_\star(z)} \end{pmatrix}. \tag{4.6}$$

The vectors $\varphi_j^\star(z)$ must satisfy $P_j^\star(z)\varphi_j^{\star\prime}(z) \equiv 0$ (see (2.22)). We compute, dropping the arguments,

$$\chi_j^{\star\prime} = \begin{pmatrix} \frac{1}{2}\sqrt{\frac{p_\star}{a_\star}} \left(\frac{a_\star}{p_\star}\right)' \\ \frac{(-1)^j}{2} \sqrt{\frac{a_\star}{p_\star}} \left(\frac{p_\star}{a_\star}\right)' - i \frac{c_\star'}{\sqrt{p_\star a_\star}} + \frac{i}{2} \frac{c_\star(p_\star a_\star)'}{(p_\star a_\star)^{3/2}} \end{pmatrix} \tag{4.7}$$

and

$$P_j^\star \chi_j^{\star\prime} = i \frac{(-1)^j}{2} \frac{c_\star a_\star' - c_\star' a_\star}{p_\star a_\star} \chi_j^\star. \tag{4.8}$$

Consequently, the vectors

$$\varphi_j^\star = \frac{e^{-i((-1)^j/2)\int_{-\infty}^{z} (c_\star a_\star' - c_\star' a_\star / p_\star a_\star)\, dz}}{\|\chi_j^\star(-\infty)\|} \chi_j^\star \tag{4.9}$$

normalized to 1 at $z = -\infty$, satisfy condition (2.22). □

COROLLARY 4.2. *Let $z_k \in X$ and let η_k be a counterclockwise-oriented loop based at the origin which encircles the disc $D(z_k, r)$ only and passes through no point of $Y_\star$. Then the quantity $e^{i\theta^\star_{12}(k)}$ defined in Lemma* 3.1 *is given by*

$$e^{i\theta^\star_{12}(k)} = -ie^{i\pi n^\star_k} \frac{\|\chi^\star_1(-\infty)\|}{\|\chi^\star_2(-\infty)\|} e^{-i/2\int_{\eta_k}(c_\star a'_\star - c'_\star a_\star/\sqrt{\rho_\star}a_\star)\,dz} e^{-2i\sigma_\star(0)},$$

where $n^\star_k \in \mathbf{Z}$ depends on $a_\star$ and η_k.

Proof. It is always possible to choose a loop η_k as described. By Lemma 3.1 we have

$$\sqrt{\rho_\star(0|\eta_k)} = e^{i\pi}\sqrt{\rho_\star(0)} \tag{4.10}$$

and

$$a_\star(0|\eta_k) = e^{i2\pi n^\star_k} a_\star(0) \tag{4.11}$$

with $n^\star_k \in \mathbf{Z}$ since $a_\star(z)$ is single valued in $\tilde{\Omega}$. As a consequence

$$\chi^\star_2(0|\eta_k) = -ie^{i\pi n^\star_k}\chi^\star_1(0). \tag{4.12}$$

Finally,

$$\sigma_\star(0|\eta_k) = \frac{1}{2}\int_{\eta_k} \frac{c_\star a'_\star - c'_\star a_\star}{\sqrt{\rho_\star}a_\star} dz + \sigma_\star(0) \tag{4.13}$$

so that

$$\varphi^\star_2(0|\eta_k) = \varphi^\star_1(0)(-i)e^{i\pi n^\star_k} \frac{\|\chi^\star_1(-\infty)\|}{\|\chi^\star_2(-\infty)\|} e^{-i/2\int_{\eta_k}(c_\star a'_\star - c'_\star a_\star/\sqrt{\rho_\star}a_\star)\,dz} e^{-2i\sigma_\star(0)}. \tag{4.14}$$

□

Consider now the two decompositions (4.1) and (4.2). The relation between the coefficients associated with the choice of eigenvectors made in Lemma 4.1 is given by the following corollary.

COROLLARY 4.3. *The coefficients $c^\star_j(\pm\infty)$ and $c_j(\pm\infty)$ defined by* (4.1), (4.2), *and Lemma* 4.1 *are such that*

$$c_j(-\infty) = c^\star_j(-\infty)e^{-i/\varepsilon\int_0^{-\infty} e^\star_j(x) - e_j(x)\,dx},$$

$$c_j(+\infty) = c^\star_j(+\infty)e^{-i(-1)^j(\sigma_\star(+\infty)-\sigma(+\infty))}e^{-i/\varepsilon\int_0^{+\infty} e^\star_j(x)-e_j(x)\,dx}$$

for $j = 1, 2$.

Proof. We write the operator A under the form

$$A(z) = \begin{pmatrix} ic(z) & a(z) \\ b(z) & -ic(z) \end{pmatrix}, \tag{4.15}$$

where we can assume, without loss of generality, that

$$\lim_{x\to\pm\infty} a(x) = a(\pm\infty) \neq 0. \tag{4.16}$$

Indeed, we can always perform a change of orthonormal basis which amounts to replacing $A(z)$ by $S^{-1}A(z)S$, where S is a constant unitary matrix. Since the gap condition holds at $\pm\infty, A(\pm\infty) \neq 0$. Thus, we can bring nonzero elements in the upper right corner of the matrices $S^{-1}A(\pm\infty)S$ by taking for S a rotation matrix in the plane of suitable angle. The corresponding eigenvectors $\varphi_j(z)$ are given by the expressions of Lemma 4.1, where the indices $\star$ are dropped. Because the operators $A(x)$ and $A_{q^\star}(x)$ coincide at $|x| = \infty$, we have

$$\chi_j^\star(\pm\infty) = \chi_j(\pm\infty) \tag{4.17}$$

and

$$\varphi_j^\star(-\infty) = \varphi_j(-\infty). \tag{4.18}$$

Hence

$$\varphi_j^\star(+\infty) = \varphi_j(+\infty)e^{-i(-1)^j(\sigma_\star(+\infty)-\sigma(+\infty))} \equiv e^{-i\beta_j^\star}\varphi_j(+\infty), \tag{4.19}$$

so that formulae (2.43) and (2.45) apply. □

4.2. Invariants. Let us consider now the following three classes of operators $A(x)$:

1)

$$A(x) = A(x)^\dagger, \qquad x \in \mathbf{R}, \tag{4.20}$$

where † denotes the adjoint.

2)

$$A(x) = \begin{pmatrix} ic(x) & a(x) \\ b(x) & -ic(x) \end{pmatrix}, \quad a(x), b(x), c(x) \in \mathbf{R}, \quad x \in \mathbf{R}. \tag{4.21}$$

3)

$$A(x) = i\begin{pmatrix} c(x) & \alpha(x) \\ \beta(x) & -c(x) \end{pmatrix}, \quad \alpha(x), \beta(x), c(x) \in \mathbf{R}, \quad x \in \mathbf{R}. \tag{4.22}$$

Note in particular that the operator $H(x)$ in equation (1.7) belongs to the first class whereas the operators in equations (1.9) and (1.11) belong to the second class. For these classes of operators there exist expressions involving the coefficients $c_j(x)$ and $c_j^\star(x)$ which are constant for all $x \in \mathbf{R}$.

LEMMA 4.4. *If $A(x)$ belongs to class* 1, 2, *or* 3, *then the operators $A_q(x)$ constructed by means of the iterative scheme* (2.11), (2.12) *belong to the same class, for any $q \leq q^\star$.*

The proof of this lemma is obtained by a straightforward induction and will therefore be omitted.

LEMMA 4.5. i) *If $A(x)$ belongs to class* 1, *then*

$$|c_1(x)|^2 + |c_2(x)|^2 = |c_1^\star(x)|^2 + |c_2^\star(x)|^2 \equiv I, \qquad x \in \mathbf{R},$$

where I is constant.

ii) *If $A(x)$ belongs to class 2 or 3, then*

$$|c_1(x)|^2 - |c_2(x)|^2 = |c_1^\star(x)|^2 - |c_2^\star(x)|^2 \equiv I, \qquad x \in \mathbf{R},$$

where I is constant.

Proof. The first assertion is a direct consequence of the fact that $U(x, x_0)$, $W(x, x_0)$, and $W_{q^\star}(x, x_0)$ are unitary if $A(x)$ and $A_{q^\star}(x)$ are self-adjoint. Assume now that $A(x)$ belongs to the second class and let

$$G = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{4.23}$$

If $\varphi(x)$ is solution of equation (1.12),

$$i\varepsilon\varphi(x)' = A(x)\varphi(x), \tag{4.24}$$

then $\overline{G\varphi(x)}$ is another solution of this equation. Indeed, $G^2 = \mathbf{1}$ so that we can write

$$i\varepsilon\overline{G\varphi(x)}' = \overline{-Gi\varepsilon\varphi(x)'} = \overline{-GA(x)GG\varphi(x)} \tag{4.25}$$

and we compute

$$\overline{-GA(x)G} = A(x), \qquad x \in \mathbf{R}. \tag{4.26}$$

Therefore, as $\mathrm{tr}A(x) \equiv 0$, the following determinant is constant for any real x:

$$\det(\varphi(x), \overline{G\varphi(x)}) = \text{ constant.} \tag{4.27}$$

Observe that the eigenvectors constructed in Lemma 4.1 satisfy the identity

$$\overline{G\varphi_j(x)} = \varphi_k(x), \qquad j \neq k \tag{4.28}$$

since $\sigma(x)$ is real and $\|\chi_j(x)\|$ is independent of $j = 1, 2$ for real $a(x), b(x)$, and $c(x)$. Then we obtain from the reality of $e_j(x)$ and $e_1(x) = -e_2(x)$ that

$$\overline{G\varphi(x)} = \overline{c_1(x)}e^{-i/\varepsilon\int_0^x e_2(x')\,dx'}\varphi_2(x) + \overline{c_2(x)}e^{-i/\varepsilon\int_0^x e_1(x')\,dx'}\varphi_1(x). \tag{4.29}$$

It remains to use the multilinearity of the determinant to get

$$\det(\varphi(x), \overline{G\varphi(x)}) = (|c_1(x)|^2 - |c_2(x)|^2)\det(\varphi_1(x), \varphi_2(x)); \tag{4.30}$$

we compute

$$\det(\varphi_1(x), \varphi_2(x)) = 2\frac{\sqrt{\rho(-\infty)}}{a(-\infty) + b(-\infty)} \tag{4.31}$$

using $\rho(x) = a(x)b(x) - (c(x))^2$. The identities (4.28) and (4.29) are also true for the eigenvectors $\varphi_j^\star(x)$ due to Lemma 4.4. Hence the same argument and (4.17) show that

$$\det(\varphi(x), \overline{G\varphi(x)}) = (|c_1^\star(x)|^2 - |c_2^\star(x)|^2)2\frac{\sqrt{\rho(-\infty)}}{a(-\infty) + b(-\infty)} = \text{constant.} \tag{4.32}$$

If $A(x)$ belongs to the third class, we proceed in a similar way. In this case, if $\varphi(x)$ is a solution of (1.12), $\overline{\varphi(x)}$ is another solution and we obtain from the explicit formulae of Lemma 4.1 (with the choice $\sqrt{i} = e^{i\pi/4}$)

$$\overline{\varphi_j(x)} = -i\varphi_k(x). \tag{4.33}$$

Finally we compute

$$\begin{aligned} \det(\varphi(x), \overline{\varphi(x)}) &= (|c_1(x)|^2 - |c_2(x)|^2) 2 \frac{\sqrt{\rho(-\infty)}}{\beta(-\infty) - \alpha(-\infty)} \\ &= (|c_1^\star(x)|^2 - |c_2^\star(x)|^2) 2 \frac{\sqrt{\rho(-\infty)}}{\beta(-\infty) - \alpha(-\infty)} = \text{constant.} \end{aligned} \tag{4.34}$$

□

Remark. It follows from (4.29) that if $(c_1(x), c_2(x))$ are solutions of (3.18), then $(\overline{c_2(x)}, \overline{c_1(x)})$ provide another solution of (3.18) when $A(x)$ belongs to class 2 or 3. The corresponding symmetry property when $A(x)$ belongs to class 1 is that if $(c_1(x), c_2(x))$ satisfy (3.18), then $(\overline{c_2(x)}, -\overline{c_1(x)})$ satisfy (3.18) as well. This property can be derived from (3.18) directly by using the antiself-adjointness of $K_q(x), q \leq q^\star$ in this case [13].

4.3. Main applications. a) Let $A(x)$ be a 2×2 hermitian matrix, $x \in \mathbf{R}$, as in equation (1.7). The equation

$$i\varepsilon \frac{d\varphi(x)}{dx} = A(x)\varphi(x), \qquad \varepsilon \to 0 \tag{4.35}$$

describes the adiabatic limit of the dynamics of a two-level quantum mechanical system. The squared modulus of the element S_{21} gives the probability $\mathcal{P}(\varepsilon)$ of a quantum transition over infinite time between the two eigenstates of the system.

COROLLARY 4.6. *If $A(x)$ is hermitian and satisfies Conditions* I–III,

$$\mathcal{P}(\varepsilon) = |S_{21}|^2 = \left| \sum_{k=1}^{p} e^{-i/\varepsilon \int_{\eta_k} e_1^\star(z,\varepsilon)\, dz} e^{-i\theta_{12}^\star(k,\varepsilon)} \right|^2 + O(e^{-\kappa\varepsilon^{-1}}) e^{\varepsilon^{-1} 2\mathrm{Im}\Delta_{12}(z_1)}.$$

b) Let $A(x)$ be the matrix (1.11)

$$A(x) = \begin{pmatrix} 0 & 1 \\ E - V(x) & 0 \end{pmatrix} \tag{4.36}$$

associated with the semiclassical regime of Schrödinger equation

$$-\varepsilon^2 \frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = E\psi(x), \qquad \varepsilon \to 0, \tag{4.37}$$

where $\inf_{x \in \mathbf{R}} E - V(x) > 0$. A solution $\varphi(x)$ of (1.11) characterized by the asymptotic conditions $c_1(-\infty) = 0, c_2(-\infty) = 1$ describes a particle coming from the right whose energy is strictly above the potential barrier $V(x)$. The reflection coefficient $\mathcal{R}(\varepsilon)$ for this scattering process is then defined by $\mathcal{R}(\varepsilon) = |\frac{c_1(+\infty)}{c_2(+\infty)}|^2$. As it stands here, it cannot be computed from the knowledge of S_{21}. However, as a consequence of Lemma 4.5 and the remark following it, we can write

$$\mathcal{R}(\varepsilon) = \frac{|\tilde{c}_2(+\infty)|^2}{1 + |\tilde{c}_2(+\infty)|^2}, \tag{4.38}$$

where $\tilde{c}_1(-\infty) = 1$ and $\tilde{c}_2(-\infty) = 0$. Hence we have the following corollary.

COROLLARY 4.7. *If $A(x)$ given by (4.46) satisfies conditions* I–III,

$$\mathcal{R}(\varepsilon) = \frac{|S_{21}|^2}{1+|S_{21}|^2} = \left| \sum_{k=1}^{p} e^{-i/\varepsilon \int_{\eta_k} e_1^*(z,\varepsilon)\,dz} e^{-i\theta_{12}^*(k,\varepsilon)} \right|^2 + O(e^{-\kappa\varepsilon^{-1}}) e^{\varepsilon^{-1} 2\mathrm{Im}\Delta_{12}(z_1)}.$$

c) Let $A(x)$ be the matrix

$$A(x) = \begin{pmatrix} 0 & 1 \\ \omega^2(x) & 0 \end{pmatrix} \tag{4.39}$$

associated with the equation of motion (1.9) of a classical oscillator whose frequency varies slowly with time

$$\varepsilon^2 \frac{d^2u(x)}{dx^2} = -\omega^2(x)u(x), \quad u(0) = u_0, \varepsilon \frac{du(0)}{dx} = u_1, \quad \varepsilon \to 0. \tag{4.40}$$

We assume that the initial values u_0 and u_1 are independent of ε. In terms of the variable $u(x)$, the adiabatic invariant (1.6) reads (keeping the same notation J)

$$J(x,\varepsilon) = \frac{\varepsilon^2|u'(x)|^2 + \omega^2(x)|u(x)|^2}{\omega(x)}. \tag{4.41}$$

Note that we do not require the initial values u_0 and u_1 to be real. Let us express $\Delta J(\varepsilon)$ in terms of the elements of the matrix S. We set

$$\Omega(x) = \begin{pmatrix} \omega(x) & 0 \\ 0 & \dfrac{1}{\omega(x)} \end{pmatrix} \tag{4.42}$$

so that we have with $\varphi(x)$ defined by (1.8)

$$J(x,\varepsilon) = \langle \varphi(x) | \Omega(x) \varphi(x) \rangle. \tag{4.43}$$

Writing

$$\varphi(x) = \sum_{j=1}^{2} d_j(x) e^{-i/\varepsilon \int_0^x (-1)^j \omega(x')\,dx'} \varphi_j(x), \tag{4.44}$$

where

$$\varphi_j(x) = \begin{pmatrix} \dfrac{1}{\sqrt{\omega(x)}} \\ (-1)^j \sqrt{\omega(x)} \end{pmatrix} \sqrt{\frac{\omega(-\infty)}{1+\omega^2(-\infty)}}, \tag{4.45}$$

we compute

$$J(x,\varepsilon) = 2\frac{\omega(-\infty)}{1+\omega^2(-\infty)} (|d_1(x)|^2 + |d_2(x)|^2). \tag{4.46}$$

Let us introduce the coefficients $d_j^\star(x)$ by

$$\varphi(x) = \sum_{j=1}^{2} d_j^\star(x) e^{-i/\varepsilon \int_0^x e_j^\star(x')\,dx'} \varphi_j^\star(x) \tag{4.47}$$

satisfying the initial condition

$$\varphi(0) = \begin{pmatrix} u_0 \\ iu_1 \end{pmatrix} = d_1^\star(0)\varphi_1^\star(0) + d_2^\star(0)\varphi_2^\star(0). \tag{4.48}$$

This last equation and Lemma 4.1 allow us to express the $d_j^\star(0)$ as functions of u_0 and u_1 and we have in particular $d_j^\star(0) = O(1)$. As a consequence of Corollary 4.3 we have $|d_j(\pm\infty)| = |d_j^\star(\pm\infty)|, j = 1, 2$, so that

$$\Delta J(\varepsilon) = 2\frac{\omega(-\infty)}{1+\omega^2(-\infty)}(|d_1^\star(+\infty)|^2 + |d_2^\star(+\infty)|^2 - |d_1^\star(-\infty)|^2 - |d_2^\star(-\infty)|^2). \tag{4.49}$$

Then it results from the linearity of equation (3.18) and from the remark following the proof of Lemma 4.5 that we can write

$$\begin{pmatrix} d_1^\star(x) \\ d_2^\star(x) \end{pmatrix} = \alpha(\varepsilon) \begin{pmatrix} c_1^\star(x) \\ c_2^\star(x) \end{pmatrix} + \beta(\varepsilon) \begin{pmatrix} \overline{c_2^\star(x)} \\ \overline{c_1^\star(x)} \end{pmatrix}, \tag{4.50}$$

where the $c_j^\star(x)$ satisfy (3.18) as well with boundary conditions $c_1^\star(-\infty) = 1, c_2^\star(-\infty) = 0$. These boundary conditions together with equation (2.46) allow us to express the constants $\alpha(\varepsilon)$ and $\beta(\varepsilon)$ as functions of the $d_j^\star(0)$ which are defined by the initial condition (4.48):

$$\begin{pmatrix} d_1^\star(-\infty) \\ d_2^\star(-\infty) \end{pmatrix} = \begin{pmatrix} \alpha(\varepsilon) \\ \beta(\varepsilon) \end{pmatrix} = \begin{pmatrix} d_1^\star(0) + O(e^{-\kappa\varepsilon^{-1}}) \\ d_2^\star(0) + O(e^{-\kappa\varepsilon^{-1}}) \end{pmatrix}. \tag{4.51}$$

We can now express the total variation of the adiabatic invariant as a function of the matrix S and the initial conditions using (4.49) and Lemma 4.5:

$$\begin{aligned} \Delta J(\varepsilon) = 2\frac{\omega(-\infty)}{1+\omega^2(-\infty)}[&4\mathrm{Re}\{\alpha(\varepsilon)\overline{\beta(\varepsilon)}c_1^\star(+\infty)c_2^\star(+\infty)\} \\ &+ 2|c_2^\star(+\infty)|^2(|\alpha(\varepsilon)|^2 + |\beta(\varepsilon)|^2)] \end{aligned} \tag{4.52}$$

Hence, by (4.51) and Corollary 4.3, we have the following corollary.

COROLLARY 4.8. *If $A(x)$ given by* (4.39) *satisfies conditions* I–III,

$$\begin{aligned} \Delta J(\varepsilon) = 2\frac{\omega(-\infty)}{1+\omega^2(-\infty)}[&4\mathrm{Re}\{d_1(-\infty)\overline{d_2(-\infty)}e^{+2i/\varepsilon \int_0^{-\infty}(e_1^\star(x,\varepsilon)\to e_1(x))\,dx} S_{11}S_{21}\} \\ &+ 2|S_{21}|^2(|d_1(-\infty)|^2 + |d_2(-\infty)|^2)]. \end{aligned}$$

If $d_1(-\infty)\overline{d_2(-\infty)} = 0$

$$\begin{aligned} \Delta J(\varepsilon) = 4\frac{\omega(-\infty)}{1+\omega^2(-\infty)} &\left|\sum_{k=1}^{p} e^{-i/\varepsilon \int_{\eta_k} e_1^\star(z,\varepsilon)\,dz} e^{-i\theta_{12}^\star(k,\varepsilon)}\right|^2 (|d_1(-\infty)|^2 + |d_2(-\infty)|^2) \\ &+ O(e^{-\kappa\varepsilon^{-1}})e^{\varepsilon^{-1}2\mathrm{Im}\Delta_{12}(z_1)}. \end{aligned}$$

If $d_1(-\infty)\overline{d_2(-\infty)} \neq 0$

$$\Delta J(\varepsilon) = 8\frac{\omega(-\infty)}{1+\omega^2(-\infty)}\mathrm{Re}\left\{d_1^\star(0)\overline{d_2^\star(0)}\sum_{k=1}^{p} e^{-i/\varepsilon\int_{\eta_k} e_1^\star(z,\varepsilon)\,dz} e^{-i\theta_{12}^\star(k,\varepsilon)}\right\} + O(e^{-\kappa\varepsilon^{-1}})e^{\varepsilon^{-1}\mathrm{Im}\Delta_{12}(z_1)},$$

where the quantities $d_j^\star(0) = O(1)$ *are determined by the initial condition* (4.48).

Remark. i) The coefficients d_j are $O(1)$ since the initial conditions u_0 and u_1 are independent of ε.

ii) The condition $d_1(-\infty)\overline{d_2(-\infty)} \neq 0$ is equivalent to $d_1(0)\overline{d_2(0)} \neq 0$. From (4.45) and (4.48) we compute

$$d_1(0) = \frac{1}{2}\sqrt{\frac{1+\omega^2(-\infty)}{\omega(-\infty)}}\left(u_0\sqrt{\omega(0)} - \frac{i}{\sqrt{\omega(0)}}u_1\right),$$

$$d_2(0) = \frac{1}{2}\sqrt{\frac{1+\omega^2(-\infty)}{\omega(-\infty)}}\left(u_0\sqrt{\omega(0)} + \frac{i}{\sqrt{\omega(0)}}u_1\right), \tag{4.53}$$

so that $d_1(-\infty)\overline{d_2(-\infty)} \neq 0$ is equivalent to $u_1 \neq \pm i\omega(0)u_0$. This condition is always true for real initial values u_0 and u_1.

Appendix. We briefly describe in this appendix an explicit example of potential $V(x)$ for which the semiclassical above barrier reflection coefficient can be computed by applying the general theory developed in this paper. Consider the potential

$$V(x) = \frac{1}{1+x^4} \tag{A.1}$$

and choose an energy level $E > 1$. Then the function

$$p^2(x) \equiv \rho(x) = E - \frac{1}{1+x^4} \tag{A.2}$$

is positive for any $x \in \mathbf{R}$. This function is meromorphic in $\mathbf{C}$ with first-order poles at the points

$$y_k = e^{i((\pi/4)+k(\pi/2))}, \qquad k = 0,1,2,3 \tag{A.3}$$

and first-order zeros at the points

$$z_k = \left(1 - \frac{1}{E}\right)^{1/4} e^{i((\pi/4)+k(\pi/2))}, \qquad k = 0,1,2,3. \tag{A.4}$$

Hence the matrix $A(x)$ given by (1.11) has an analytic continuation in the set $\Omega \equiv \mathbf{C}\setminus\{y_1,y_2,y_3,y_4\}$. The Stokes lines are obtained by studying the level lines of the multivalued function $\int_0^z dz'p(z')$ in the set Ω. By a numerical study, we see that these lines behave in the first quadrant of the complex plane as described in Fig. 7.

We can show by exploiting the symmetries of the function ρ that these lines are symmetric with respect to both the real and imaginary axes. Hence, Conditions I, II, and III are satisfied and the above barrier reflection coefficient can be computed

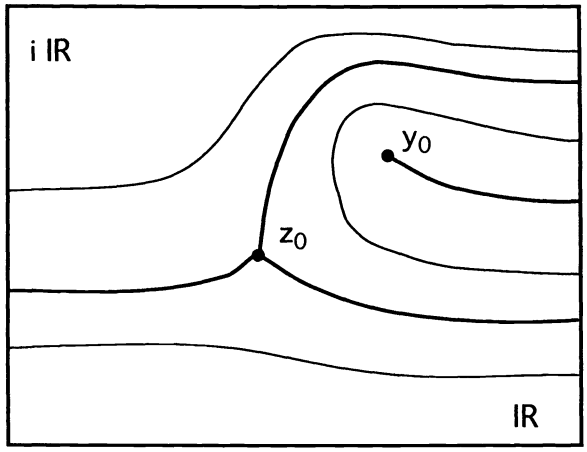


FIG. 7. *The level lines of* $\int_0^z dz' p(z')$ *in the first quadrant of the complex plane.*

asymptotically as $\hbar$ goes to zero using the method explained above. In particular, we see from Corollary 4.2 that in the first-order asymptotic formula, $\theta_{12}(k), k = 0, 1$ is real since the function $c(z) \equiv 0$ and $\|\chi_1(\pm\infty)\| = \|\chi_2(\pm\infty)\|$; see (4.7). Hence it remains to compute $\int_{\eta_k} p(z)\, dz, k = 1, 2$, to get the first-order asymptotic formula for $\mathcal{R}(\hbar)$. Moreover, the presence of two first-order zeros in the upper half-plane linked by a Stokes line shows that an interference phenomenon takes place (Stückelberg oscillations) at the first order already, even though the potential barrier displays one bump only. The high-order corrections can be systematically computed using the theory developed in this paper; we omit this computational aspect here.

Acknowledgments. We thank the referee for constructive suggestions and C. Ballif for computing numerically the Stokes lines of the example in the appendix.

REFERENCES

[1] N. FRÖMAN AND P. O. FRÖMAN, *JWKB Approximation, Contributions to the Theory*, North Holland, Amsterdam, 1965.

[2] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley Intersciences, New York, 1965.

[3] M. V. FEDORIUK, *Méthodes Asymptotiques pour les Equations Différentielles Ordinaires Linéaires*, Mir Moscou, 1987.

[4] PH. A. MARTIN AND G. NENCIU, *Semi-classical inelastic S-matrix for one-dimensional N-state systems*, Proc. Leuven Conference on the Three Levels, Micro-, Meso-, and Macro-Approaches in Physics, M. Fannes, C. Meas, and A. Verbeure, eds., Plenum Press, New York, 1994.

[5] J. DAVIS AND P. PECHUKAS, *Nonadiabatic transitions induced by a time-dependent Hamiltonian in the semiclassical/adiabatic limit: The two-state case*, J. Chem. Phys., 64 (1976), pp. 3129–3137.

[6] J.-T. HWANG AND P. PECHUKAS, *The adiabatic theorem in the complex plane and the semi-classical calculation of non-adiabatic transition amplitudes*, J. Chem. Phys., 67 (1977), pp. 4640–4653.

[7] A. JOYE, H. KUNZ, AND C.-E. PFISTER, *Exponential decay and geometric aspect of transition probabilities in the adiabatic limit*, Ann. Phys., 208 (1991), pp. 299–332.

[8] M. V. BERRY, *Geometric Amplitude Factors in Adiabatic Quantum Transitions*, Proc. Roy. Soc. London Ser. A, 430 (1990), pp. 405–411.

[9] R. E. MEYER, *Adiabatic variation, Part* II. *Action change for the simple oscillator*, J. Appl. Math. Phys., 24 (1973), pp. 517–524.

[10] W. WASOW, *Calculation of an adiabatic invariant by turning point theory*, SIAM J. Math. Anal., 5 (1974), pp. 673–700.

[11] M. V. FEDORIUK, *An adiabatic invariant for a system of linear oscillators and scattering theory*, Differential Equations, 12 (1976), pp. 713–718.

[12] J. B. KELLER AND Y. MU, *Changes in adiabatic invariants*, Ann. Phys., 205 (1991), pp. 219–227.

[13] A. JOYE, G. MILETI, AND C.-E. PFISTER, *Interferences in adiabatic transition probabilities mediated by Stokes lines*, Phys. Rev. A, 44 (1991), pp. 4280–4295.

[14] N. FRÖMAN, *Connection formulas for certain higher order phase-integral approximations*, Ann. Phys., 61 (1970), pp. 451–464.

[15] B. LUNDBORG, *Phase-integral treatment of wave reflection by real potentials*, Math. Proc. Cambridge Philos. Soc., 85 (1979), pp. 493–522.

[16] A. JOYE AND C.-E. PFISTER, *Full asymptotic expansion of transition probabilities in the adiabatic limit*, J. Phys. A, 24 (1991), pp. 753–766.

[17] S. G. KREIN, *Linear Differential Equations in Banach Spaces*, American Mathematical Society, Providence, RI, 1971.

[18] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Heidelberg, New York, 1980.

[19] A. JOYE AND C.-E. PFISTER, *Superadiabatic evolution and adiabatic transition probability between two non-degenerate levels isolated in the spectrum*, J. Math. Phys., 34 (1993), pp. 454–479.

[20] G. NENCIU, *Linear adiabatic theory: Exponential estimates*, Comm. Math. Phys., 152 (1993), pp. 479–496.

[21] A. MARTINEZ, *Precise exponential estimates in adiabatic theory*, Université Paris-Nord, 1993, preprint.

[22] J. SJÖSTRAND, *Projecteurs adiabatiques du point de vue pseudodifférentiel*, C.R. Acad. Sci. Paris Sér. I Math, 317 (1993), pp. 217–220.

[23] A. JOYE AND C.-E. PFISTER, *Exponentially small adiabatic invariant for the Schrödinger equation*, Comm. Math. Phys., 140 (1991), pp. 15–41.

[24] A. JOYE, *Non-trivial prefactors in adiabatic transition probabilities induced by high order complex degeneracies*, J. Phys. A, 26 (1993), pp. 6517–6540.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 978–998, July 1995

009

BIFURCATION OF SPATIAL CENTRAL CONFIGURATIONS FROM PLANAR ONES*

RICHARD MOECKEL† AND CARLES SIMÓ‡

Abstract. Central configurations are important special solutions of the Newtonian N-body problem of celestial mechanics. In this paper a highly symmetrical case is studied. As the masses are varied, spatial central configurations appear through bifurcation from planar ones. In particular, spatial configurations can be found which are arbitrarily close to being planar.

Key words. celestial mechanics, central configurations, bifurcation

AMS subject classifications. 34, 58, 70, 85

1. Introduction.

This paper is concerned with certain highly symmetrical central configurations of the Newtonian N-body problem. It is shown that under certain conditions, spatial central configurations (that is, nonplanar ones) bifurcate from planar ones as the masses are varied. Moreover, the bifurcation is described in detail.

Recall that the N-body problem concerns the motion of N point particles with masses $m_j \in \mathbf{R}^+$ and positions $q_j \in \mathbf{R}^3$, where $j = 1, \ldots, N$. The motion is governed by Newton's law

$$m_j \ddot{q}_j = \frac{\partial U}{\partial q_j}.$$

Here $U(q)$ is the Newtonian potential

$$U(q) = \sum_{\substack{(i,j) \\ i<j}} \frac{m_i m_j}{|q_i - q_j|}.$$

Let $q = (q_1, \ldots, q_N)^T \in \mathbf{R}^{3N}$ and $M = \operatorname{diag}(m_1, m_1, m_1, \ldots, m_N, m_N, m_N)$. Then the equation of motion can be written as follows:

$$\ddot{q} = M^{-1} \frac{\partial U}{\partial q}.$$

In studying this problem, there is no loss of generality in assuming that the center of mass of the particles is at the origin: $m_1 q_1 + \cdots + m_N q_N = 0$. Because the potential is singular when two particles have the same position, it is natural to assume that the configuration avoids the set $\Delta = \{q : q_i = q_j \text{ for some } i \neq j\}$.

DEFINITION 1. *A configuration $q \in \mathbf{R}^3 \backslash \Delta$ is called a* central configuration *if there is some constant λ such that*

$$M^{-1} \frac{\partial U}{\partial q} = \lambda q.$$

* Received by the editors May 3, 1993; accepted for publication (in revised form) November 29, 1993.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. This research was supported by the National Science Foundation, the Sloan Foundation, and the Eidgenösische Technische Hochschule, Zürich.

‡ Departament de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain. This research was partially supported by CICYT grant ESP91-403.

Using Newton's law one sees that $\ddot{q}_j = \lambda q_j$; in other words, the acceleration vector of every particle is pointing toward the origin with the magnitude of the acceleration proportional to the distance from the origin. From the fact that the potential is homogeneous of degree -1, it is easy to show that $\lambda = -U(q)/(q^T M q)$.

If the masses are released from a central configuration with zero initial velocity, the configuration collapses homothetically to the origin. This explains the use of the term central. In the planar case one can choose the initial velocities to produce circular and elliptical periodic orbits with the configuration remaining similar to the initial central configuration. Central configurations are important for celestial mechanics, but here they will be studied for their own sake.

Central configurations are the rest points of a certain gradient flow. Introduce a metric on $\mathbf{R}^{3N}$: $\langle q, q\rangle = q^T M q$ and let

$$S = \{q : \langle q, q\rangle = 1, m_1 q_1 + \cdots + m_N q_N = 0\}$$

denote the unit sphere with respect to this metric in the subspace where the center of mass is at the origin. Since the subspace has dimension $3N - 3$, S is a $(3N - 4)$-dimensional sphere. It follows from the homogeneity of the Newtonian potential that the vector field $X = M^{-1}\frac{\partial U}{\partial q} + \lambda q$, where $\lambda = U(q)$, is tangent to S. Moreover, it has rest points exactly at the central configurations with $\langle q, q\rangle = 1$. Finally, it is easy to check that $\langle X(q), v\rangle = DU(q)v$ for every $q \in S$ and $v \in T_q S$. These facts show that the vector field $M^{-1}\frac{\partial U}{\partial q} + U(q)q$ is the gradient of $U|_S$, the restriction of $U(q)$ to the unit sphere S with respect to the metric $\langle \cdot, \cdot \rangle$. Moreover, the rest points of this vector field are exactly the central configurations in S. Note that any central configuration is homothetic to one in S. Thus the problem of finding central configurations is essentially that of finding rest points of the gradient flow of $U|_S$, or, alternatively, of finding critical points of $U|_S$.

The gradient flow preserves certain submanifolds of S. For example, the set of all collinear configurations and the set of all planar configurations are invariant. In addition, some sets of configurations with symmetry are preserved. One of these is the main object of study in this work.

2. Symmetrical configurations. Consider the set Σ of all configurations in $\mathbf{R}^3$ consisting of two regular N-gons, $N \geq 2$, lying in horizontal planes, centered on a common vertical axis, and aligned so that corresponding vertices lie in the same vertical half-plane (see Fig. 1). Suppose that all of the particles in one N-gon have mass μ and all of the particles in the other N-gon have mass 1. Then, it follows from symmetry that the gradient vector field of $U|_S$ is tangent to Σ, so Σ is invariant under the flow. To find central configurations in Σ, it suffices to study the gradient flow of $U|_\Sigma$. As Σ is only two-dimensional, this is a great simplification.

Figure 1 shows three parameters (r, s, t) which can be used to describe such a configuration. The metric in these coordinates is

$$\langle q, q\rangle = N\mu r^2 + N s^2 + N\frac{\mu}{1+\mu} t^2.$$

To express the potential in these coordinates, note that the potential of a regular N-gon of unit size and unit masses is simply NA, where

$$A = \frac{1}{2}\sum_{j=1}^{N-1} \frac{1}{|1 - e^{i2\pi j/N}|} = \frac{1}{4}\sum_{j=1}^{N-1} \csc(\pi j/N).$$

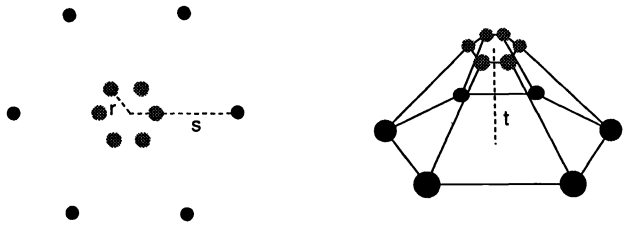


FIG. 1. *The symmetrical configurations considered here.*

The potentials of the two N-gons in the configuration are simply $\frac{NA\mu^2}{r}$ and $\frac{NA}{s}$. In addition, there will be the potential arising from the interaction of the two N-gons. The total potential is

$$U(q) = N\left(\frac{\mu^2}{r} + \frac{1}{s}\right)A + N\mu\sum_{j=1}^{N}\frac{1}{\Delta_j},$$

where

$$\Delta_j^2 = |r - se^{i2\pi j/N}|^2 + t^2 = r^2 + s^2 + t^2 - 2rs\cos(2\pi j/N).$$

The factors of N in both the metric and the potential can be dropped without essentially changing the results. Then the unit sphere Σ is given by the equation

$$\mu r^2 + s^2 + \frac{\mu}{1+\mu}t^2 = 1.$$

The gradient of $U|_\Sigma$ with respect to this metric is

$$\left(\frac{1}{\mu}U_r, U_s, \frac{1+\mu}{\mu}U_t\right) + (Ur, Us, Ut).$$

Thus we obtain the following differential equations for the gradient flow:

$$\begin{aligned}
\dot{r} &= -\frac{\mu}{r^2}A - Fr + Gs + Ur,\\
\dot{s} &= -\frac{1}{s^2}A - \mu Fs + \mu Gr + Us,\\
\dot{t} &= -(1+\mu)Ft + Ut,
\end{aligned}$$

where

$$F = \sum_{j=1}^{N}\frac{1}{\Delta_j^3}$$

and

$$G = \sum_{j=1}^{N}\frac{\cos(2\pi j/N)}{\Delta_j^3}.$$

The central configurations in Σ are obtained by setting all three derivatives equal to zero; however, since the vector field is tangent to the unit sphere, only two of these equations are independent.

For the computations that follow, it is convenient to introduce two parameters on the unit sphere. Let $x = \frac{r}{s}$ and $y = \frac{t}{s}$. Then, a short computation gives $\dot{x}$ and $\dot{y}$. However, the equations can be simplified by multiplication by s^3. This can be viewed as a change of time scale in the gradient flow and will not affect the results. Denoting the derivative with respect to this new time variable by a prime instead of a dot, the equations become

$$\begin{aligned} x' &= f(x,y) - xg(x,y), \\ y' &= -yg(x,y), \end{aligned} \tag{1}$$

where

$$\begin{aligned} f(x,y) &= \mu x P + Q - \frac{\mu}{x^2} A, \\ g(x,y) &= P + \mu x Q - A, \end{aligned}$$

and

$$\begin{aligned} P = s^3 F &= \sum_{j=1}^{N} \frac{1}{d_j^3}, \\ Q = s^3 G &= \sum_{j=1}^{N} \frac{\cos(2\pi j/N)}{d_j^3}, \\ d_j^2 = s^{-2}\Delta_j^2 &= 1 + x^2 + y^2 - 2x\cos(2\pi j/N). \end{aligned}$$

The equations for central configurations are $x' = y' = 0$. These are invariant under the transformation $(x, y, \mu) \to (\frac{1}{x}, \frac{y}{x}, \frac{1}{\mu})$ together with a time rescaling by μx^3. Thus, it suffices to find all central configurations with $0 < x \leq 1$. In the next section, it will be shown that there is always a unique planar central configuration with x in this range.

In what follows, the value of A as a function of N will play an important role. The following lemma gives an asymptotic expansion for $A(N)$.

LEMMA 1. *Let* $A = \frac{1}{4}\sum_{j=1}^{N-1} \csc(\pi j/N)$. *Then,* $A(N)$ *has the following asymptotic expansion for* N *large:*

$$A(N) \sim \frac{N}{2\pi}\left(\gamma + \log\frac{2N}{\pi}\right) + \sum_{k\geq 1} \frac{(-1)^k(2^{2k-1}-1)B_{2k}^2\pi^{2k-1}}{(2k)(2k)!}\frac{1}{N^{2k-1}}, \tag{2}$$

where γ *stands for the Euler–Mascheroni constant and* B_{2k} *stands for the Bernoulli numbers.*

Proof. Consider the case where N is even; the odd case is similar. Then

$$A = \frac{1}{4}\left(2\sum_{j=1}^{N/2-1} \csc(\pi j/N) + 1\right). \tag{3}$$

The cosecant can be written as follows [1]:

$$\begin{aligned} &\csc(z) = z^{-1} + h(z), \qquad \text{where} \\ &h(z) = \sum_{k\geq 1} \frac{(-1)^{k-1}2(2^{2k-1}-1)B_{2k}z^{2k-1}}{(2k)!}. \end{aligned} \tag{4}$$

Here $h(z)$ is analytic for $|z| < \pi$. In our case, $|z| = |\frac{\pi j}{N}| < \frac{\pi}{2}$ for all j.

To carry out the sum in (2) we split the cosecant as in (4). The first part can be summed using the digamma function (the logarithmic derivative of the Γ function)

$$\psi(m) = -\gamma + \sum_{j=1}^{m-1} j^{-1},$$

which has the asymptotic expansion

$$\psi(z) \sim \log(z) - \frac{1}{2z} - \sum_{k\geq 1} \frac{B_{2k}}{2k\, z^{2k}}. \tag{5}$$

Recall the Euler–MacLaurin summation formula

$$\begin{aligned}\sum_{j=1}^{m-1} g(j) \sim \int_0^m g(z)\, dz - \frac{1}{2}\left(g(0) + g(m)\right) \\ + \sum_{k\geq 1} \frac{B_{2k}}{(2k)!}\left(g^{(2k-1)}(m) - g^{(2k-1)}(0)\right).\end{aligned} \tag{6}$$

This will be applied to $g(z) = h(\pi z/N)$. We use the following facts, which follow from the definition of $g(z)$:

$$\int_0^{N/2} \left(\csc(z\pi/N) - (z\pi/N)^{-1}\right) dz = \frac{N}{\pi}\log\frac{4}{\pi}, \qquad g(0) = 0, \quad g\left(\frac{\pi}{2}\right) = 1 - \frac{2}{\pi},$$

$$g^{(2k-1)}(0) = \frac{(-1)^{k-1} 2(2^{2k-1}-1) B_{2k}}{2k} \left(\frac{\pi}{N}\right)^{2k-1},$$

$$\begin{aligned} g^{(2k-1)}\left(\frac{N}{2}\right) &= \left(\frac{\pi}{N}\right)^{2k-1} \frac{d^{2k-1}}{dz^{2k-1}}(\csc(z) - z^{-1})|_{z=\pi/2} \\ &= \frac{2^{2k}}{\pi}(2k-1)!\frac{1}{N^{2k-1}}. \end{aligned}$$

Substituting these into (6) and using (5) we obtain

$$\begin{aligned}\sum_{j=1}^{N/2-1} \csc(\pi j/N) \sim \frac{N}{\pi}\left(\gamma + \log\frac{N}{2} - \frac{1}{N} - \sum_{k\geq 1}\frac{B_{2k}}{2k\,(N/2)^{2k}}\right) + \frac{N}{\pi}\log\frac{4}{\pi} - \frac{1}{2}\left(1 - \frac{2}{\pi}\right) \\ + \sum_{k\geq 1}\frac{B_{2k}}{(2k)!}\left(\frac{2(2k-1)!}{\pi\,(N/2)^{2k-1}} - \frac{(-1)^{k-1}2(2^{2k-1}-1)B_{2k}}{2k}\left(\frac{\pi}{N}\right)^{2k-1}\right).\end{aligned}$$

Finally, using this in (3) leads, after simplification, to the desired result. □

For most purposes, it is enough to use a few terms of the asymptotic expansion (2). Recall that $\gamma \approx 0.57721566490153286$, and that the first few Bernoulli numbers are $B_0 = 1$, $B_1 = \frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_4 = -\frac{1}{30}$, $B_6 = \frac{1}{42}$. So

$$A \approx \frac{N}{2\pi}\left(\gamma + \log\frac{2N}{\pi}\right) - \frac{\pi}{144N} + \frac{7\pi^3}{86400N^3} - \frac{31\pi^5}{7620480N^5}. \tag{7}$$

More generally, the Bernoulli numbers are defined by

$$\frac{t}{e^t - 1} = \sum_{m\geq 0}\frac{B_m t^m}{m!},$$

and satisfy $B_{2n-1} = 0$ for $n \geq 1$ and $B_{2n} = (-1)^{n-1}2(2n)!\zeta(2n)/(2\pi)^{2n}$, where $\zeta(s) = \sum_{k\geq 1} k^{-s}$ is the Riemann zeta function. It is easy to check, using the Euler–MacLaurin summation formula with remainder, that the absolute error is less than the absolute value of the first neglected term. The best estimate is obtained by taking terms up to $k_0 = [\pi n]$ or $k_0 = [\pi n] + 1$, where [ ] denotes the integer part. With this choice, the absolute error is less than $N^{1/2}\exp(-2\pi N)$. Even for $N = 2$, when $A = \frac{1}{4}$, the choice $k_0 = 6$ gives an error less than 1.5×10^{-6}. Furthermore, for practical computations, formula (7) gives a relative error less than 10^{-16} for $N \geq 47$. With $k_0 = 18$, one could achieve this error for $N \geq 6$.

3. The planar case. The set of planar configurations consisting of two nested regular N-gons is obtained from the above family by setting $y = 0$. The restriction $0 < x \leq 1$ means that the polygon with particles of mass μ is inside the other one. At $x = 0$ the inner polygon shrinks to a point, while at $x = 1$ the two polygons collide. In both cases, the potential becomes infinite. It will be shown that there is a unique intermediate value x_p representing a central configuration. When $y = 0$, the second equation in (1) is true and the first will determine x_p.

Let $h(x) = f(x,0) - xg(x,0)$. Then $\lim_{x\to 0} h(x) = -\infty$, while $\lim_{x\to 1} h(x) = +\infty$. Thus, to prove the existence of a unique zero of $h(x)$, it suffices to show that $h(x)$ is increasing. Now $h(x)$ can be written as follows:

$$h(x) = \left(x - \frac{\mu}{x^2}\right) A + \mu x \left(P(x,0) - xQ(x,0)\right) + \left(Q(x,0) - xP(x,0)\right).$$

Let

$$\phi(x) = \sum_{j=1}^{N} \frac{1}{d_j}.$$

It follows from the definitions that

$$\phi(x) = \left(1 + x^2\right) P(x,0) - 2xQ(x,0),$$

which implies

$$P(x,0) - xQ(x,0) = \phi(x) + x\left(Q(x,0) - xP(x,0)\right).$$

Moreover, the expression $(Q(x,0) - xP(x,0))$ is simply $\frac{d\phi}{dx}$. Substitution into the formula for $h(x)$ gives

$$h(x) = \left(x - \frac{\mu}{x^2}\right) A + \mu x\phi(x) + \left(1 + \mu x^2\right) \frac{d\phi}{dx}. \tag{8}$$

The first term is clearly increasing. The other two terms are handled by the following lemma, which will also be useful later.

LEMMA 2. *Let $\phi_\lambda(x) = \sum_{j=1}^{N} 1/d_j^\lambda$, where $\lambda > 0$. Then, for $0 < x < 1$, $\phi_\lambda(x)$ and all of its derivatives are positive. Moreover, the same is true for $\psi_\lambda(x) = \sum_{j=1}^{N} \cos(2\pi j/N)/d_j^\lambda$.*

Proof. An explicit power series expansion for ϕ_λ will be obtained, and it will turn out that all of the coefficients are nonnegative. Recall the expansion

$$\frac{1}{(1-z)^{\lambda/2}} = \sum_{n=0}^{\infty} \alpha_n z^n,$$

where the coefficients are positive numbers depending on λ. Following Euler, one can write

$$\frac{1}{d_j^\lambda} = \frac{1}{(1 - xe^{+i2\pi j/N})^{\lambda/2}} \frac{1}{(1 - xe^{-i2\pi j/N})^{\lambda/2}}.$$

Using the power series above gives

$$\frac{1}{d_j^\lambda} = \sum_{n=0}^{\infty} c_n^j x^n,$$

where

$$c_n^j = \sum_{k+l=n} \alpha_k \alpha_l e^{i2\pi j(k-l)/N}.$$

This must be summed for $j = 1, \ldots, N$. But the sum over the roots of unity is N if $k \equiv l \bmod N$, and 0 otherwise. Thus

$$\phi_\lambda(x) = N \sum_{n=0}^{\infty} \left(\sum_{\substack{k+l=n \\ k \equiv l \bmod N}} \alpha_k \alpha_l \right) x^n.$$

Here, all of the coefficients are nonnegative and infinitely many (including all of the even ones) are positive. The claim about $\phi_\lambda(x)$ and its derivatives follows. A similar computation works for $\psi_\lambda(x)$. □

For future reference, note that $P(x, 0) = \phi_3(x)$ and $Q(x, 0) = \psi_3(x)$, so it follows that both of them are increasing for $0 < x < 1$. It also follows that $\phi_\lambda(x)$, $\psi_\lambda(x)$, and their derivatives are negative for $x > 1$.

From (8) and Lemma 2 with $\lambda = 1$, one finds for any mass ratio μ exactly two planar central configurations of nested regular N-gons, one $0 < x_p(\mu) < 1$, and one $\tilde{x}_p(\mu) > 1$. By symmetry we have $\tilde{x}_p(\mu) = 1/x_p(\frac{1}{\mu})$. Before turning to the nonplanar case it is interesting to investigate the dependence of x_p on the parameters μ and N; only the solution between 0 and 1 will be considered. From the definition, $h(x_p(\mu)) = 0$. Differentiating this gives

$$h_\mu(x_p(\mu)) + h_x(x_p(\mu)) \frac{dx_p(\mu)}{d\mu} = 0.$$

It has already been shown that $h_x > 0$. On the other hand, equation (8) gives

$$0 = x_p A + \frac{d\phi}{dx}(x_p) + \mu h_\mu(x_p),$$

which shows that $h_\mu(x_p(\mu)) < 0$. Thus

$$\frac{dx_p(\mu)}{d\mu} > 0.$$

In other words, as the mass of the inner polygon is increased, the size of that polygon in the central configuration also increases.

In the limit as $\mu \to 0$, $x_p(\mu) \to 0$. To see this, note that the only negative contribution to (8) is the term $-A\mu/x^2$. Using the power series in Lemma 2, one finds the more precise asymptotics

$$\mu = \frac{2A + N}{2A} x_p(\mu)^3 + O(x_p^4).$$

On the other hand, in the limit as $\mu \to \infty$, $x_p(\mu) \to x_p(\infty) < 1$, where $x_p(\infty)$ is the solution of

$$-\frac{A}{x^2} + x\phi(x) + x^2\frac{d\phi}{dx} = 0.$$

The case of equal masses is of special interest as well. Taking $\mu = 1$, the equation for the central configuration simplifies to

$$(1 + x + x^2)\, A = x^2(1 + x)Q(x, 0). \tag{9}$$

Finally, one can view the symmetrical central configuration determined by $x_p(\mu)$ as a rest point of the gradient flow on Σ. It will be important in what follows to know the eigenvalues of the linearized flow near this rest point. These linearized equations are quite simple since the derivative matrix of the vector field (1) is diagonal when $y = 0$. Thus, the eigenvectors are parallel to the x and y axes with eigenvalues

$$\lambda_p = h_x(x_p(\mu)),$$
$$\lambda_\perp = -g(x_p(\mu), 0).$$

Here, the subscripts indicate whether the eigenvector is planar or perpendicular to the planar manifold. It has been noted several times already that $h_x > 0$, so the eigenvalue within the planar manifold is always positive. It turns out that while $\lambda_\perp$ is usually negative, it can be positive under certain conditions. This fact was already noted in [3], but a proof will now be given.

First, note that $\lambda_\perp(\mu) = -g(x_p(\mu), 0)$ is a decreasing function of μ. To see this, just differentiate

$$-\frac{d\lambda_\perp}{d\mu} = g_x(x_p(\mu), 0)\frac{dx_p}{d\mu} + g_\mu(x_p(\mu), 0).$$

Now, $g_x(x, 0) = P_x(x, 0) + \mu Q_x(x, 0) + \mu Q(x, 0)$, and the remarks following Lemma 2 show that this is positive. Also, it was just shown that $dx_p/d\mu > 0$. Finally, $g_\mu(x, 0) = xQ(x, 0) > 0$.

For the equal mass case, the eigenvalue is negative. In fact, using (9) to eliminate A in the formula for $g(x_p(1), 0)$ yields

$$-\lambda_\perp(1) = P(x_p(1), 0) + \frac{x_p(1)}{1 + x_p(1) + x_p(1)^2}Q(x_p(1), 0) > 0.$$

Since $\lambda_\perp(\mu)$ is decreasing, it must also be negative for $\mu > 1$. On the other hand, as $\mu \to 0$, $x_p(\mu) \to 0$ as well, so

$$\lim_{\mu \to 0} \lambda_\perp(\mu) = A - P(0, 0) = A - N.$$

Thus, there will be an interval on which $\lambda_\perp(\mu) > 0$ if and only if $A > N$.

Lemma 1 gave an asymptotic expansion for $A(N)$. The first term in the expansion, together with the remark about the error after the proof, shows that $\frac{A(N)}{N}$ is increasing. Thus we only need to find the first N such that $\frac{A(N)}{N} > 1$. Solving $A(x) = x$ using just the first term of the expansion gives $x = \frac{\pi}{2}\exp(2\pi - \gamma) \approx 472.27$. This suggests $N = 473$. According to the first term,

$$\frac{A(472)}{472} \approx 0.99990875, \qquad \text{while} \qquad \frac{A(473)}{473} \approx 1.00024558,$$

and the first neglected term has absolute value less than 10^{-7}. Thus $A > N$ if and only if $N \geq 473$.

If $N < 473$, the eigenvalue $\lambda_\perp(\mu)$ is negative for all values of μ. But if $N \geq 473$, there is a unique bifurcation value $\mu_0(N)$ such that $\lambda_\perp(\mu) > 0$ for $\mu < \mu_0$ and $\lambda_\perp(\mu) < 0$ for $\mu > \mu_0$. The results of this section can be summarized as follows.

THEOREM 1. *For every mass ratio μ, there are exactly two planar central configurations consisting of two nested regular N-gons. For one of these, the ratio x_p of the sizes of the two polygons is less than 1, and for the other it is greater than 1. For most values of the parameters N and μ, these represent saddle points of the gradient flow of $U|_\Sigma$; however, for $N \geq 473$ there is a constant $\mu_0(N) < 1$ such that for $\mu < \mu_0$ and $\mu > 1/\mu_0$, the central configuration with the smaller masses on the inner polygon is a repeller.*

It follows that in the case $N \geq 473$, nonplanar central configurations must bifurcate from the planar central configuration as μ increases through μ_0, a conclusion already drawn in [3]. In the next section, the exact nature of this bifurcation will become clear.

4. The nonplanar case. In this section, all of the central configurations with $y \neq 0$ will be found. In this case, setting the vector field (1) to zero gives

$$f(x, y) = g(x, y) = 0. \tag{10}$$

Since only y^2 appears in the equations, the solutions will come in pairs related by reflection through a horizontal plane. The main result is that there is either exactly one such pair or none at all, depending on whether the eigenvalue $\lambda_\perp(\mu)$ is negative or positive.

Some a priori information on the shape of a possible nonplanar central configuration will be useful. When $\mu = 1$, it is clear that there will be a pair of central configurations with the two N-gons stacked one above the other, that is, with $x = 1$. It will now be shown that for $\mu < 1$, any nonplanar central configuration has $x < 1$. For $\mu > 1$ the situation is reversed. In other words, the N-gon with the smaller masses is smaller. To see this, consider the quantity $f - \mu g$, which must vanish for a nonplanar central configuration. From the definitions one finds

$$f - \mu g = \mu(x - 1)P + (1 - \mu^2 x)Q + \mu A\left(1 - \frac{1}{x^2}\right).$$

Suppose that $\mu \leq 1$ but that, in contradiction to the claim, $x > 1$. Then the fact that $P > Q$ gives $\mu(x - 1)P > \mu(x - 1)Q$, so

$$f - \mu g > \mu(1 - \mu)xQ + \mu A\left(1 - \frac{1}{x^2}\right) > 0.$$

This proves the claim and, incidentally, also shows that for equal masses, the stacked N-gons are the only possibility. With this information in hand, there is no loss of generality in considering the cases $\mu \leq 1$ and $x \leq 1$.

The solutions of (10) with $y > 0$ will be found by first fixing x and solving $g(x, y) = 0$ for $y(x)$, and then solving $f(x, y(x)) = 0$ for x. Note that the partial derivative $g_y(x, y) < 0$ when $y > 0$. Furthermore, $\lim_{y \to \infty} g(x, y) = -A < 0$. Thus, for a given value of x, there will be a unique solution of $g(x, y) = 0$ provided only that $g(x, 0) \geq 0$. The behavior of $g(x, 0)$ was studied in the last section. It increases from $N - A$ to ∞ as x goes from 0 to 1. Thus, if $N < 473$ it is always positive, whereas if

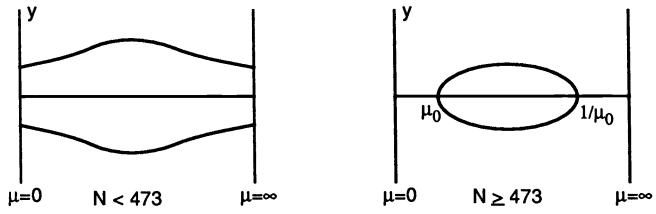


FIG. 2. *Bifurcation diagrams.*

$N \geq 473$ there will be a unique $x_0(\mu) \in (0,1)$ at which point it changes from negative to positive. Thus, in the former case one can solve for $y(x) > 0$ for any $x \in (0,1]$, while in the latter one can find $y(x) > 0$ only for $x \in (x_0, 1]$ and $y(x_0) = 0$.

It is important to know the relationship between the endpoint $x_0(\mu)$ and the position of the planar central configuration $x_p(\mu)$. Recall that the value of $g(x_p(\mu), 0)$ is just minus the eigenvalue $\lambda_\perp(\mu)$. Thus the sign of $g(x_p(\mu), 0)$ is negative if $\mu < \mu_0$ and positive if $\mu > \mu_0$, where $\mu_0 = \mu_0(N)$ of Theorem 1. On the other hand, at $x_0(\mu)$ one has $g(x_0(\mu), 0) = 0$ by definition. Since $g(x, 0)$ is increasing, it follows that the order will be $x_p(\mu) < x_0(\mu)$ if $\mu < \mu_0$ and $x_0(\mu) < x_p(\mu)$ if $\mu > \mu_0$.

Once $y(x)$ has been found, the equation $f(x, y(x)) = 0$ must be satisfied. It will be demonstrated below that $\frac{df(x,y(x))}{dx} > 0$. This already implies that there is at most one solution. Now, we must check the values at the endpoints. It is always the case that $f(1, y(1)) \geq 0$ with equality only for $\mu = 1$. To see this, note that when $x = 1$, $f = \mu P + Q - \mu A$. But from the definition of $y(1)$, $P(1, y(1)) + \mu Q(1, y(1)) = A$. Using this to eliminate A in the formula for f gives $f(1, y(1)) = (1-\mu^2)Q(1, y(1)) > 0$. Thus everything depends on the sign of $f(x, y(x))$ at the other endpoint.

In the case in which $N < 473$, $\lim_{x \to 0} f(x, y(x)) = -\infty$. In the case in which $N \geq 473$, the sign of $f(x_0, y(x_0)) = f(x_0, 0)$ is negative if and only if $\mu > \mu_0$. To see this, recall that if $\mu > \mu_0$ then $x_0(\mu) < x_p(\mu)$. Also, the increasing function $h(x) = f(x, 0) - xg(x, 0)$ vanishes at x_p by definition. Thus $h(x_0) < 0$, so $f(x_0, 0) < x_0 g(x_0, 0) = 0$ as desired. If $\mu < \mu_0$ the inequalities are reversed.

Thus, if either $N < 473$ or $N \geq 473$, and $\mu > \mu_0$, there will be a unique pair of spatial central configurations $(x_s(\mu), \pm y_s(\mu))$; otherwise there are no nonplanar central configurations. These results are summarized in the bifurcation diagrams of Fig. 2 which show the y coordinates of the central configurations as functions of μ. The bifurcation whose existence was inferred in §3 is now seen to be a simple pitchfork bifurcation in which nonplanar central configurations bifurcate from the planar central configuration with the small masses inside.

As already remarked in [3], there are several surprising consequences of this phenomenon. First, nonplanar central configurations can be almost planar; this is in contrast to the fact that any noncollinear central configuration is actually far from collinear. Second, in the case where no nonplanar central configurations exist, the absolute minimum potential energy (within the symmetric configurations considered) is attained at a planar configuration. This is surprising because, intuitively, there is more room for the masses to spread out in three dimensions. Finally, again in the case where no nonplanar central configurations exist, the fact that the eigenvalue $\lambda_\perp$ is positive implies the existence of certain interesting total collision orbits in the $2N$ body problem. Briefly, the relation between central configurations, and total collision orbits is as follows: along any orbit which experiences such a collision, the configura-

tion (rescaled to unit size) tends to a central configuration, and the set of all orbits converging to a given central configuration is locally a submanifold of the phase space [2]. The fact that $\lambda_\perp > 0$ implies that the manifold of orbits converging to the planar central configuration contains nonplanar orbits. In other words, there are total collision orbits which are initially nonplanar but which are asymptotically planar. Once again, this is in contrast to the collinear case where any orbit which is asymptotically collinear must, in fact, be collinear for all time.

We still must prove the inequality $\frac{df(x,y(x))}{dx} > 0$. Some computation yields the formula

$$\begin{aligned} g_y \frac{df(x,y(x))}{dx} &= f_x g_y - f_y g_x \\ &= \frac{2A\mu}{x^3} g_y + \mu(PP_y - QQ_y) \\ &\quad + \mu^2 x(PQ_y - QP_y) + (1-\mu^2x^2)(Q_xP_y - P_xQ_y). \end{aligned}$$

Since $g_y < 0$, it suffices to show that all terms on the right side are negative. This is clearly true of the first term. The second term is also simple to handle. For the third term write

$$PQ_y - QP_y = -3y \sum_{j,k=1}^{N} \frac{1}{d_j^3 d_k^3} (\cos(2\pi j/N) - \cos(2\pi k/N)) \left(\frac{1}{d_j^2} - \frac{1}{d_k^2} \right).$$

It is clear that $\cos(2\pi j/N) > \cos(2\pi k/N)$ if and only if $d_k^2 > d_j^2$, so every term with $j \neq k$ in the sum is positive. Finally, in the fourth term in the expression, the factor $1-\mu^2x^2$ is positive, since it is assumed that μ and x are less than 1. The other factor is

$$Q_xP_y - P_xQ_y = -9y \left[\sum_{j=1}^{N} \frac{1}{d_j^5} \cdot \sum_{j=1}^{N} \frac{\cos^2(2\pi j/N)}{d_j^5} - \left(\sum_{j=1}^{N} \frac{\cos(2\pi j/N)}{d_j^5} \right)^2 \right].$$

The sums in the square brackets can be interpreted as inner products of the vectors $u = (1, \ldots, 1)$ and $v = (\cos(2\pi/N), \ldots, \cos(2\pi N/N))$ with respect to a metric with coefficients d_j^{-5}. Then the square bracket is just $|u|^2|v|^2 - (u,v)^2$, which is positive by the Schwarz inequality. This completes the proof.

The behavior of the spatial central configuration $(x_s(\mu), y_s(\mu))$ as μ varies is quite simple. Since the family begins by bifurcating from the planar central configuration, one has $(x_s(\mu_0), y_s(\mu_0)) = (x_p(\mu_0), 0)$. At equal masses the spatial configuration is stacked N-gons so $(x_s(\mu), y_s(\mu)) = (1, y_s(1))$, where $y_s(1)$ satisfies

$$g(1,y) = P(1,y) + Q(1,y) - A = 0.$$

For $\mu_0 < \mu < 1$, both $x_s(\mu)$ and $y_s(\mu)$ are strictly increasing. To see this, differentiate the equation $f(x_s, y_s) = g(x_s, y_s) = 0$ to obtain

$$\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} \begin{bmatrix} \frac{dx_s}{d\mu} \\ \frac{dy_s}{d\mu} \end{bmatrix} = - \begin{bmatrix} f_\mu \\ g_\mu \end{bmatrix}.$$

Solving this with Cramer's rule and using the inequalities $f_x, g_x, g_\mu > 0$ and $f_y, g_y, f_\mu < 0$, which are easily checked, shows that both x_s and y_s are increasing (recall that the determinant has already been shown to be negative).

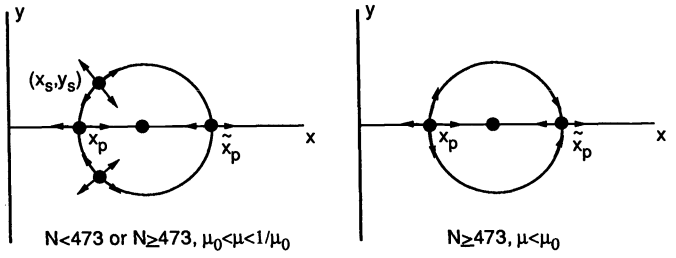


FIG. 3. *Phase portraits for the gradient flow.*

Another remark about the shape of the spatial configurations is that $x_p(\mu) < x_s(\mu) < \tilde{x}_p(\mu)$ always. By symmetry, it suffices to show the first inequality. This can be done by considering the function $h(x,y) = f(x,y) - xg(x,y)$. This is just x' in equation (1), so it vanishes at the central configurations. In particular, $h(x_p,0) = 0$. It can be shown that this implies $h(x_p,y) < 0$ for $y > 0$. Taking $y = y(x_p)$ as in the proof above shows that $f(x_p,y(x_p)) < x_p g(x_p,y(x_p)) = 0$. But $f(x,y(x))$ is increasing and reaches 0 at (x_s,y_s). Thus $x_p < x_s$.

Next, the eigenvalues of the linearized differential equations near a nonplanar central configuration will be studied. Since it is a gradient flow, both eigenvalues are real. Furthermore, when the nonplanar configurations exist, they are the absolute minima of $U|_\Sigma$, since the planar configurations are saddle points. Now the determinant of the linearization of (1) at a nonplanar central configuration is $y(f_y g_x - f_x g_y)$. This expression has already been shown to be positive. This shows that, in fact, the nonplanar configurations are nondegenerate minima, so both of the eigenvalues of the linearization are positive.

Combining this with the information about the planar central configuration from §3 leads to the following phase portraits of the gradient flow on $U|_\Sigma$ shown in Fig. 3. In these pictures, the point $(1,0)$, the y-axis, and the unbounded regions of the half-plane are all attracting since the potential becomes infinite there.

The main results about spatial central configurations can be summarized as follows.

THEOREM 2. *If $N < 473$, there is a unique pair of spatial central configurations of parallel regular N-gons. If $N \geq 473$, there are no such central configurations for $\mu < \mu_0(N)$. At $\mu = \mu_0$ a unique pair bifurcates from the planar central configuration with the smaller masses on the inner polygon. This remains the unique pair of spatial central configurations until $\mu = 1/\mu_0$, where a similar bifurcation occurs in reverse, so that for $\mu > 1/\mu_0$, only the planar central configurations remain. When the spatial central configurations exist, they are repellers in the gradient flow of $U|_\Sigma$.*

5. Asymptotics and some numerical results. Finally, we give additional information on the asymptotic behavior of the solutions, both planar and spatial, for large N and other limiting cases. We also display the evolution of the solutions, when changing μ, for several values of N, comparing the numerical results with the asymptotic ones. Since both μ and N will be varied, we will adopt the notation $x_p(\mu,N)$ for the planar central configuration in $(0,1)$ and $(x_s(\mu,N), y_s(\mu,N))$ for the spatial central configuration with $y > 0$.

First we shall consider the asymptotic behavior when μ goes to 0. We need to introduce several auxiliary variables. Let $c_j = \cos(2\pi j/N)$ and $s_k = \frac{1}{N}\sum_{j=1}^{N} c_j^k$. We

immediately obtain (recall $N \geq 2$) the following:

$$
\begin{aligned}
&s_0 = 1, \quad s_1 = 0, \\
&s_2 = 1 \quad \text{if } N = 2, \quad \tfrac{1}{2} \text{ otherwise}, \\
&s_3 = \tfrac{1}{4} \quad \text{if } N = 3, \quad 0 \text{ otherwise}, \\
&s_4 = 1 \quad \text{if } N = 2, \quad \tfrac{1}{2} \text{ if } N = 4, \quad \tfrac{3}{8} \text{ otherwise}.
\end{aligned} \tag{11}
$$

We need the expansions of $P(x,y)$, $Q(x,y)$ ($P(x,0)$, $Q(x,0)$ for the planar case) near $x = 0$. Let $\Delta = (1+y^2)^{1/2}$. For the planar case $\Delta = 1$. Then the derivatives of P and Q with respect to x are easily obtained as follows:

$$
\begin{aligned}
&P(0,y) = \frac{N}{\Delta^3}, \quad P_x(0,y) = 0, \quad P_{xx}(0,y) = \left(-\frac{3}{\Delta^5} + \frac{15s_2}{\Delta^7}\right) N, \\
&Q(0,y) = 0, \quad Q_x(0,y) = \frac{3s_2 N}{\Delta^5}, \quad Q_{xx}(0.y) = \frac{15s_3 N}{\Delta^7}, \\
&Q_{xxx}(0,y) = \left(-\frac{45s_2}{\Delta^7} + \frac{105s_4}{\Delta^9}\right) N,
\end{aligned} \tag{12}
$$

where the values of s_k are given in (11).

In the planar case we have

$$\mu x^3 P(x,0) + x^2 Q(x,0) - \mu A = x^3 P(x,0) + \mu x^4 Q(x,0) - Ax^3.$$

Hence, using (12),

$$\mu O(x^3) + x^3 \left[3s_2 + A - 1 - \frac{15}{2} s_3 x + \left(\frac{35}{2} s_4 - 15s_2 + \frac{3}{2}\right) x^2 + O(x^3)\right] = \mu \frac{A}{N}. \tag{13}$$

From (13) we get the following proposition.

PROPOSITION 1. *For μ small, the values of $x_p(\mu, N)$ have the following expansion:*

$$x_p(\mu, N) = e_1 \mu^{1/3} + e_2 \mu^{2/3} + e_3 \mu + O(\mu^{4/3}), \tag{14}$$

where

$$e_1 = \left(\frac{A}{\psi N}\right)^{1/3}, \quad e_2 = -\frac{5}{2}\frac{s_3}{\psi} e_1^2, \quad e_3 = \left(\frac{75 s_3^2}{4\psi^2} - \frac{35s_4 - 30s_2 + 2}{6\psi}\right) e_1^3, \tag{15}$$

where $\psi = 3s_2 - 1 + \frac{A}{N}$.

It follows from (14), (15), and Lemma 1, that this is a good approximation provided $\mu \log N$ is small. We want to remark that the coefficient of $\mu^{1/3}$ in $x_p(\mu, N)$ increases as a function of N and tends to 1 as $N \to \infty$. The e_2 coefficient is 0 unless $N = 3$, and $e_3 \to 0$ as $N \to \infty$. The curves $x_p(\mu, N)$ increase with respect to N for all μ (see Fig. 4). This follows for large N from Proposition 4. For values up to $N = 1000$, it has been checked directly.

Let us pass to the spatial case. We know that to have solutions for μ small requires $N < 473$.

PROPOSITION 2. *In the spatial case, for μ small and $N < 473$, the values of $x_s(\mu, N)$ and $y_s(\mu, N)$ have the expansions*

$$
\begin{aligned}
x_s(\mu, N) &= \left(\frac{1}{3s_2}\left(\frac{A}{N}\right)^{2/3} \mu\right)^{1/3} + R_x, \\
y_s(\mu, N) &= y^0 + y^1 \mu^{2/3} + R_y,
\end{aligned}
$$

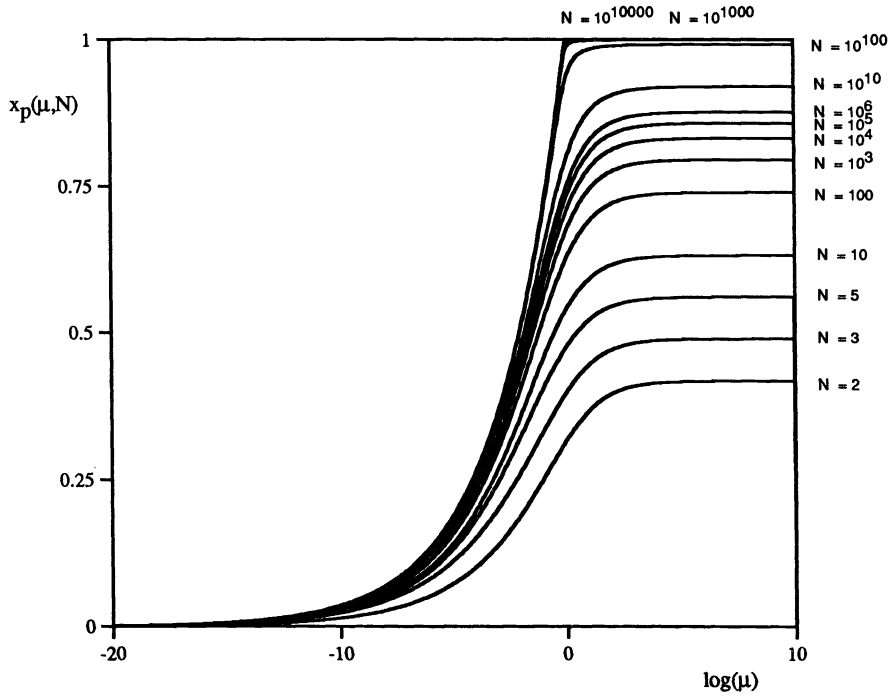


FIG. 4. *Planar case:* x_p *versus* $\log\mu$. *Solutions of* $h = 0$ *up to* $N = 1000$; *solutions of asymptotic equation* (23) $= x^3$(24), *skipping* $\mathcal{R}_1$ *and* $\mathcal{R}_3$, *for* $N > 1000$.

where

$$y^0 = \left(\left(\frac{N}{A}\right)^{2/3} - 1\right)^{1/2}, \quad y^1 = \frac{1}{2y^0}\left(\frac{5s_2}{1+(y^0)^2} - 1\right)\left(\frac{1}{3s_2}\left(\frac{A}{N}\right)^{2/3}\right)^{2/3},$$

and

$$R_x = O(\mu^{2/3}) \text{ for } N = 3, \ O(\mu) \text{ otherwise},$$

$$R_y = O(\mu) \quad \text{for } N = 3, \ O(\mu^{4/3}) \text{ otherwise}.$$

The proof is elementary from the equations

$$\begin{aligned} \mu x^3 P(x,y) + x^2 Q(x,y) - \mu A = 0, \\ P(x,y) + \mu x Q(x,y) - A = 0, \end{aligned} \tag{16}$$

and the expansions provided by (12).

Now let us proceed to predict the value of the bifurcation parameter $\mu_0(N)$ for N larger than, but close to, 472. For $N \geq 473$, let $\delta_N = \frac{A}{N} - 1$ (so $\delta_{473} \simeq 2.455 \cdot 10^{-4}$, $\delta_{474} \simeq 5.816 \cdot 10^{-4}$, $\delta_{475} \simeq 9.170 \cdot 10^{-4}$, etc.). We need the functions P, Q and the x derivatives at $(0,0)$ as given by (12), but the exceptional cases of (11) do not occur now. We also need $P_{xxxx}(0,0) = (45 - 630s_2 + 945s_4)N = \frac{675}{8}N$. From this we have the following proposition.

PROPOSITION 3. *For N close to 472 and $N > 472$, the bifurcation parameter is*

$$\mu_0(N) = \frac{4}{9}\delta_N^{3/2} - \frac{29}{54}\delta_N^{5/2} + O(\delta_N^3), \tag{17}$$

and the corresponding value of x is given by

$$x_p(\mu_0(N), N) = x_s(\mu_0(N), N) = \frac{2}{3}\delta_N^{1/2} - \frac{25}{108}\delta_N^{3/2} + O(\delta_N^2), \tag{18}$$

where $\delta_N = \frac{A}{N} - 1$.

Even for $N = 475$ the formulas (17) and (18) give relative errors less than 10^{-5}.

Next we proceed to study the behavior for N large. To this end we want to replace the sums defining P and Q by integrals with controlled error.

For $l = 1, 3$ let

$$F_l(z) = \left(1 + x^2 + y^2 - 2x\cos(2\pi z/N)\right)^{-l/2} = ((1+x)^2 + y^2 - 4x\cos^2(\pi z/N))^{-l/2}.$$

Then

$$P(x,y) = \sum_{k=1}^{N} F_3(k) = \frac{2N}{\pi}((1+x)^2 + y^2)^{-3/2} S_3, \tag{19}$$

where

$$\begin{aligned} S_3 &= \frac{\pi}{2N}\sum_{k=1}^{N}(1 - m\cos^2(\pi k/N))^{-3/2} = \int_0^{\pi/2}(1 - m\cos^2 u)^{-3/2}\,du + \mathcal{R}_3 \\ &= \frac{1}{1-m}E(m) + \mathcal{R}_3. \end{aligned} \tag{20}$$

In (20), $m = 4x((1+x)^2 + y^2)^{-1}$, $E(m)$ stands for the complete elliptic integral of second kind and parameter m, $\int_0^{\pi/2}(1 - m\cos^2 u)^{1/2}\,du$ (see [1]), and $\mathcal{R}_3$ is the remainder to be bounded below.

In a similar way,

$$(1 + x^2 + y^2)P(x,y) - 2xQ(x,y) = \sum_{k=1}^{N} F_1(k) = \frac{2N}{\pi}((1+x)^2 + y^2)^{-1/2} S_1, \tag{21}$$

where

$$\begin{aligned} S_1 &= \frac{\pi}{2N}\sum_{k=1}^{N}(1 - m\cos^2(\pi k/N))^{-1/2} = \int_0^{\pi/2}(1 - m\cos^2 u)^{-1/2}\,du + \mathcal{R}_1 \\ &= K(m) + \mathcal{R}_1, \end{aligned} \tag{22}$$

$K(m)$ being the complete elliptic integral of the first kind and parameter m, and $\mathcal{R}_1$ being the corresponding remainder.

Using (19) to (22) and substituting in the expressions of f and g we get

$$(2\mu x^3 + xw)\left(E(m) + \frac{u}{v}\mathcal{R}_3\right) - xu(K(m) + \mathcal{R}_1) = \frac{\mu u v^{1/2}\Gamma}{2}, \tag{23}$$

$$(24) \qquad (2+\mu w)\left(E(m)+\frac{u}{v}\mathcal{R}_3\right)-\mu u(K(m)+\mathcal{R}_1)=\frac{uv^{1/2}\Gamma}{2},$$

where $u=(1-x)^2+y^2$, $v=(1+x)^2+y^2$, $w=1+x^2+y^2$, $\Gamma=\frac{2A\pi}{N}$, and, then, $m=1-\frac{u}{v}$. We recall, for further use, that $E(m)$ (resp., $K(m)$) decreases (resp., increases) monotonically from $\frac{\pi}{2}$ to 1 (resp., from $\frac{\pi}{2}$ to ∞) when m goes from 0 to 1. Furthermore, for m close to 1, one has $E(m)\simeq 1+(m_1/4)\log((16/e)/m_1)$, $K(m)\simeq\frac{1}{2}\log(16/m_1)$, where $m_1=1-m$, the complementary parameter of the elliptic integrals.

To bound $\mathcal{R}_l$, $l=1,3$, we shall need rough a priori bounds of m_1 as a function of Γ.

LEMMA 3. *Let* $\Gamma=\frac{2A\pi}{N}$ *and* $m_1=((1-x)^2+y^2)/((1+x)^2+y^2)$, *where* (x,y) *are the coordinates of planar or spatial central configurations. Then, for* $N>691$, *one has*

$$m_1>\frac{0.9}{N\Gamma}.$$

Proof. First we consider the planar case ($y=0$). As $x_p(\mu,N)$ is an increasing function of μ, and m_1 decreases as a function of x for $0<x<1$, it is enough to consider the limit behavior as μ goes to infinity.

The limit equation is $P-xQ=A/x^3$. Using (19), (21), and the definition of m_1, this equation is written as follows:

$$m_1^{1/2}S_3+S_1=\frac{(1+x)\,x^{-3}\Gamma}{2}.$$

Neglecting the S_1 term and using just $k=N$ in the sum defining S_3, we have $m_1>x^3(1+x)^{-1}\pi/(N\Gamma)$. If $x<\frac{9}{11}$ we have $m_1>0.01$. If $x>\frac{9}{11}$ then $m_1>\frac{0.9}{(N\Gamma)}$. Hence in all cases we have $m_1>\min(0.01,\frac{0.9}{(N\Gamma)})$ and the minimum is the second term for $N>26$.

In the spatial case we need a bound on y. As both x and y have been proved to be increasing with respect to μ, we consider $\mu=1$ (and then $x=1$). The equations reduce to $P+Q=A$, and now $m_1=y^2/(4+y^2)$.

We claim $y^2\le 2$ if N is large enough. Assume this is not true. Then $m_1>\frac{1}{3}$. Again using (19) and (21), the equation $P+Q=A$ becomes

$$S_3-S_1=\frac{(4+y^2)^{1/2}\,\Gamma}{2}.$$

The terms in S_3 are not more than the one corresponding to $k=N$, and this one is bounded by $3^{3/2}$. Hence

$$\frac{3^{3/2}\pi}{2}=\frac{\pi}{2N}\sum_{k=1}^{N}3^{3/2}>S_3>S_3-S_1=\frac{(4+y^2)^{1/2}\Gamma}{2}>\frac{6^{1/2}\Gamma}{2}.$$

We obtain $\Gamma<3\pi 2^{-1/2}$, which is false for $N>691$. This proves the claim.

In the general spatial case ($\mu<1$), the equation $g=0$ gives $P<A$. Using the term $k=N$ in (19), we have $((1+x)^2+y^2)^{-3/2}m_1^{-3/2}<A$, that is, $m_1>A^{-2/3}/6$, where $x\le 1$ and $y^2\le 2$ have been used. This bound is larger than the bound in the statement of the lemma for $N>3$. □

Now we proceed to bound the remainders $\mathcal{R}_1$ and $\mathcal{R}_3$.

LEMMA 4. *For $N > 691$, the remainders $\mathcal{R}_3$ and $\mathcal{R}_1$ in (20) and (22) are bounded by*

$$|\mathcal{R}_l| < 6\left(\frac{N\Gamma}{0.9}\right)^l \exp(-(0.9N/\Gamma)^{1/2}). \tag{25}$$

Proof. Let us consider the functions $\varphi_l(u) = (1-m\cos^2 u)^{-l/2}$, $l = 1, 3$, appearing in (20) and (22). These are real analytic π-periodic functions with Fourier expansions $\sum_{k\in\mathbf{Z}} c_k^l e^{2iku}$. The contribution of the c_0^l term to S_l is the part expressed by the integral. Hence the remainder is bounded by

$$|\mathcal{R}_l| \le \frac{\pi}{2N} \sum_{k\in\mathbf{Z}\setminus\{0\}} |c_k^l| \cdot \left|\sum_{j=1}^{N} e^{2ikj/N}\right| = \frac{\pi}{2} \sum_{k\in N\mathbf{Z}\setminus\{0\}} |c_k^l|. \tag{26}$$

Let us bound $|c_k^l|$. Consider the functions φ_l, $l = 1, 3$, in the complex strip $|\mathrm{Im}\, u| \le m_1^{1/2}$. The maximum modulus of φ_l on that strip is attained at $u = \pm i m_1^{1/2}$ (mod π) and equals

$$\left[\frac{1 + m_1 - (1 - m_1)\cosh(2m_1^{1/2})}{2}\right]^{-l/2}.$$

Expanding around $m_1 = 0$, the square bracket is of the form $\frac{2}{3}m_1^2 + \cdots$, with all the coefficients positive. Hence

$$|\varphi_l|_{m_1^{1/2}} \le \frac{(3/2)^{l/2}}{m_1^l},$$

where $|\ |_\gamma$ denotes the supremum norm for $|\mathrm{Im}\, u| \le \gamma$. Therefore,

$$|c_k^l| \le \left(\frac{3}{2}\right)^{l/2} m_1^{-l} \exp(-|k| m_1^{1/2}).$$

By substitution in (26) we get

$$|\mathcal{R}_l| \le \frac{\pi(3/2)^{l/2}}{m_1^l} \frac{\exp(-Nm_1^{1/2})}{1 - \exp(-Nm_1^{1/2})} < \frac{6}{m_1^l} \exp(-Nm_1^{1/2}).$$

From this the lemma follows. □

We note that using the bound (25), the remainders are less than $2 \cdot 10^{-10}$ for $N \ge 50000$. Hence the remainders can be completely neglected in (20) and (22) for large N.

Remark. When we know the remainders are small, we can go back and redo Lemma 3 to have better estimates on m_1 by using formulas (23) and (24). Then we return to Lemma 4 for improved estimates on the remainders. In this way, we reach for $\mathcal{R}_l$ bounds of the form $\exp(-\frac{cN}{\Gamma})$ for some $c > 0$. Finally, the remainders can be neglected even for moderate values of N (say $N \ge 1000$).

One can solve (23) and (24), for N large, with $\frac{A\pi}{N}$ computed from Lemma 1 and skipping $\mathcal{R}_l$, $l = 1, 3$, avoiding the sums. The results are shown in Fig. 5 for $N \ge 1000$,

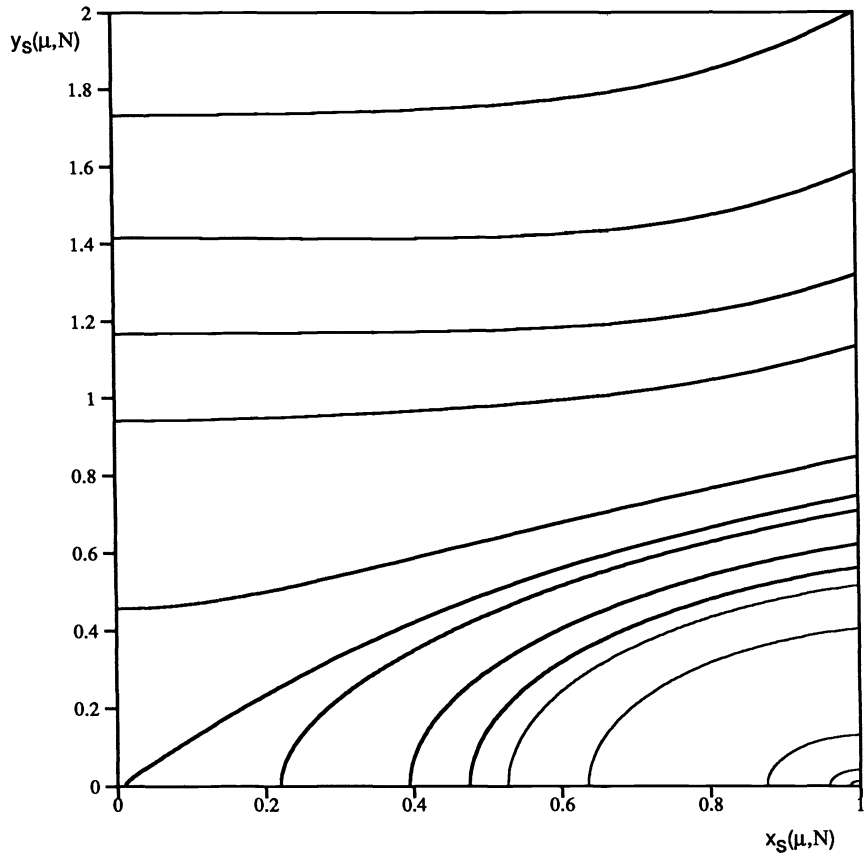


FIG. 5. *Spatial case*: y_s *versus* x_s. *Solutions of* $f = g = 0$ *for* $N < 1000$; *solutions of equations* (23) *and* (24) *for* $N \geq 1000$.

where we also display the numerical results obtained by using the original equations for $N \leq 1000$. Different curves correspond to different values of N. The variable y_s is shown against x_s. One can see that there is no intersection of the curves obtained for different N. The formulas (23) and (24) allow us to compute for very large values of N (such as 10^{10^6}). Figure 6 shows a magnification of Fig. 5 near $(x, y) = (1, 0)$ for N very large, making apparent the fact that the curves tend to circles; Fig. 7 shows a magnification near $(x, y) = (0, 0)$.

From (23) and (24) it is clear that, for any finite value of μ in $(0, 1]$, when N goes to ∞, u should go to 0.

First consider the planar case.

PROPOSITION 4. *For* N *large enough, in the planar problem* μ *can be expressed as a function of* x_p *by*

$$\mu = \frac{x\left((E(m)/(1-x)) + (K(m)/(1+x))\right) + x^3(\Gamma/2)}{(\Gamma/2) - x^3\left((E(m)/(1-x)) + (K(m)/(1+x))\right)}, \tag{27}$$

where $\Gamma = \frac{2A\pi}{N}$.

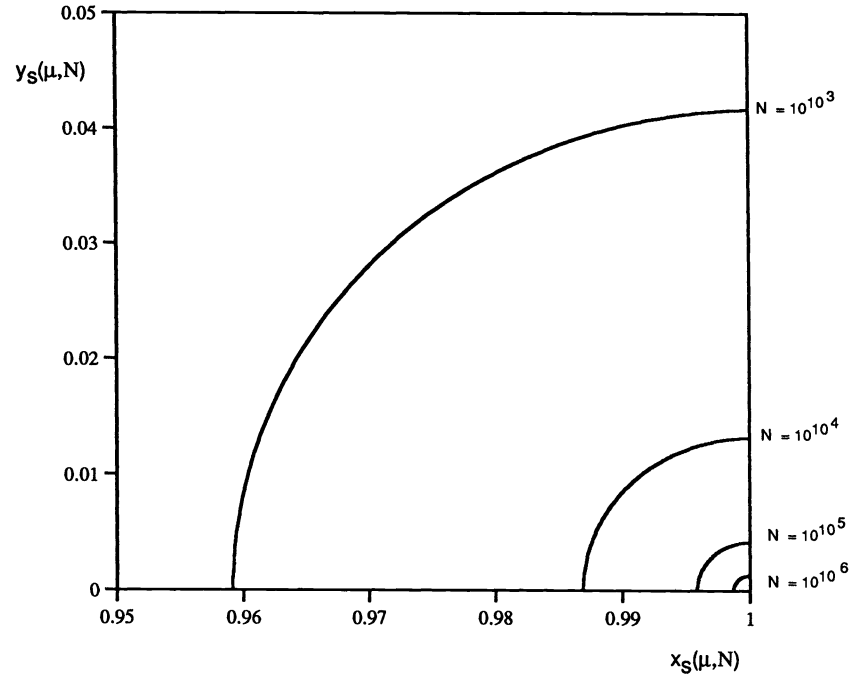


FIG. 6. *Magnification of Fig.* 5.

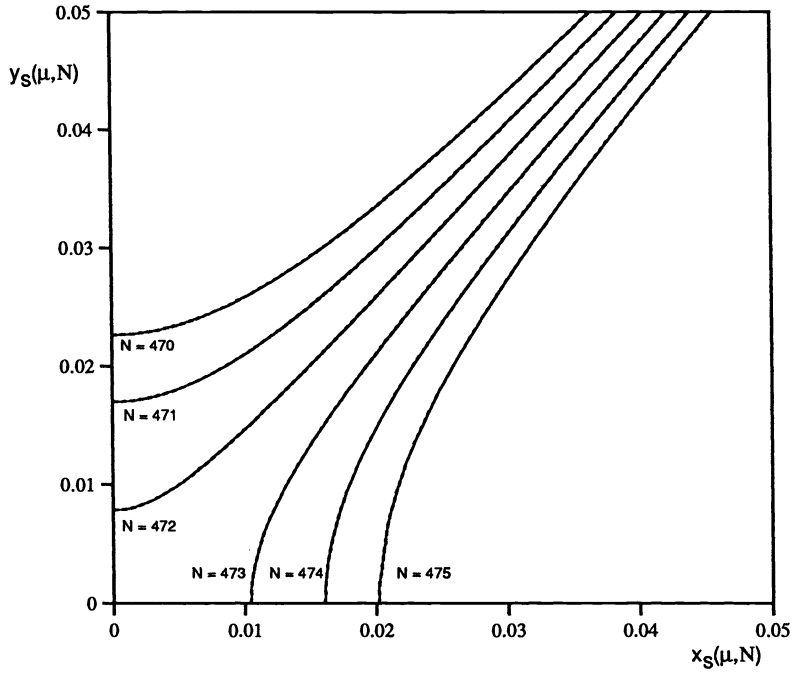


FIG. 7. *Magnification of Fig.* 5.

Furthermore, the bifurcation parameter satisfies

$$\mu_0(N) = 1 - \frac{4}{\Gamma^{1/2}} + O\left(\frac{\log \Gamma}{\Gamma}\right).$$

The value of x at the bifurcation parameter is given by

$$x_p(\mu_0(N), N) = x_s(\mu_0(N), N) = 1 - \frac{2}{\Gamma^{1/2}} + \frac{2}{\Gamma} + O\left(\frac{\log \Gamma}{\Gamma^{3/2}}\right).$$

Proof. The first part follows from $f(x,0) - xg(x,0) = 0$ by using (20) and (22) and skipping the $\mathcal{R}_l$ terms. Using $f = 0$ and $g = 0$ simultaneously, one obtains the remaining exressions. □

COROLLARY 1. *For any value of $\mu < 1$ in the planar case, one has $x_p(\mu, N) \to \mu^{1/3}$ as $N \to \infty$. For $\mu = 1$ we get $x_p(1, N) = 1 - (\frac{4}{3\Gamma})^{1/2} + O(\frac{\log \Gamma}{\Gamma})$ as $N \to \infty$. For $\mu \to \infty$, the limit satisfies $x_p(\infty, N) = 1 - \frac{2}{\Gamma} + O(\log \Gamma / \Gamma^2)$ as $N \to \infty$.*

The proof follows immediately from (27).

To close, we consider the spatial case.

PROPOSITION 5. *For any fixed $\mu \in (0, 1]$, when N goes to ∞, the spatial solution behaves like*

$$(1 - x_s)^2 + y_s^2 = \frac{4}{\Gamma} + O(\Gamma^{-3/2}). \tag{28}$$

Furthemore, a point (x_s, y_s) in the solution curve for a given N is related to the parameter μ by means of

$$1 - x_s = \frac{1 - \mu}{2} + O(\Gamma^{-1}). \tag{29}$$

For $\mu = 1$, $x_s = 1$, one can obtain a more precise expression for y as

$$y_s = \frac{2}{\Gamma^{1/2}} + \frac{O(\log \Gamma)}{\Gamma^{3/2}}. \tag{30}$$

The proof is easily obtained by introducing $\bar{z} = (1-x)/\Gamma^{1/2}$, $\bar{y} = y/\Gamma^{1/2}$, $\mu = 1 - \bar{\nu}/\Gamma^{1/2}$, inserting in (23) and (24), cancelling terms, and simplifying. One immediately gets $\bar{z}^2 + \bar{y}^2 = 4$ at order zero in Γ and, hence, (28). By equating terms in $\Gamma^{-1/2}$, one obtains the relation between $\bar{z}$ and $\bar{\nu}$ and, therefore, (29). The lack of terms in Γ^{-1} in (30) follows from the fact that only even powers of y appear.

Note that the results of Propositions 4 and 5 give the asymptotic behavior of several quantities when N goes to infinity. However, the remainders in these expressions go to zero slowly. For instance, as Γ is of the order of $\log N$, the value $\frac{\log \Gamma}{\Gamma} \simeq \frac{\log(\log N)}{\log N}$, appearing in the error term of $\mu_0(N)$, is just 0.136 for $N = 10^{10}$ and $3.36 \cdot 10^{-3}$ for N as large as 10^{1000}. Some tedious but elementary computations can give more terms in all the estimates of these propositions.

Acknowledgments. The authors are indebted to the Geometry Center in Minneapolis for use of their computing facilities during preparation of this paper and for their support of a visit of the second author. They also thank the referee for many valuable suggestions for improving the paper.

REFERENCES

[1] M. ABRAMOWITZ AND I. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965.
[2] R. MCGEHEE, *Singularities in classical celestial mechanics*, In Proc. Internat. Congress of Mathematicians, Helsinki, 1978, pp. 827–834.
[3] R. MOECKEL, *On central configurations*, Math. Z., 205 (1990), pp. 499–517.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 999–1017, July 1995

010

A MATHEMATICAL MODEL OF TRAFFIC FLOW ON A NETWORK OF UNIDIRECTIONAL ROADS*

HELGE HOLDEN† AND NILS HENRIK RISEBRO‡

Abstract. We introduce a model that describes heavy traffic on a network of unidirectional roads. The model consists of a system of initial-boundary value problems for nonlinear conservation laws. We uniquely formulate and solve the Riemann problem for such a system and, based on this, then show existence of a solution to the Cauchy problem.

Key words. traffic flow, conservation laws, network

AMS subject classification. 35L65

A new problem, which has arisen in the twentieth century, is how to organize road traffic so that the full benefits of our increased mobility can be enjoyed at the lowest cost in human life and capital. The problem has many sides—constructional, legal, educational, administrative.
— M. J. Lighthill and G. B. Whitham (1955)

Introduction. The modeling of traffic flow by conservation laws was studied by Lighthill and Whitham [18], [19], and Richards [22]. They argued as follows: Consider a road with one lane with heavy traffic so that a continuum description is a good approximation. Let $\rho = \rho(x,t)$ denote the density of cars. Then, conservation of the number of cars yields [17], [24] ("traffic hydrodynamics")

$$\rho_t + (v\rho)_x = 0, \tag{0.1}$$

where $v = v(x,t)$ denotes the velocity of cars at (x,t).

Equation (0.1) is valid for continuum descriptions of any conserved quantity. The first fundamental additional assumption we will make for the problem of traffic flow is that the velocity field v is a function only of the density ρ, $v = v(\rho)$.

We also assume certain reasonable properties of the function v: when the density is small, there is a maximum velocity $v_{\max}$, and when ρ increases to some maximum capacity $\rho_{\max}$, the velocity vanishes.

A linear interpolation then gives the simplest possible velocity

$$v = v_{\max}\left(1 - \frac{\rho}{\rho_{\max}}\right). \tag{0.2}$$

One of the earliest velocity fields to be studied was the so-called Greenberg model with

$$v(\rho) = v_0 \ln\left(\frac{\rho_{\max}}{\rho}\right) \tag{0.3}$$

supported by experimental data from the Lincoln tunnel in New York [8]. (Although v is divergent at low density, we see that the relevant quantity in (0.1) is ρv, which

* Received by the editors September 26, 1993; accepted for publication November 3, 1993.
† Department of Mathematical Sciences, The Norwegian Institute of Technology, University of Trondheim, N-7034 Trondheim, Norway (holden@imf.unit.no).
‡ Department of Mathematics, University of Oslo, P.O. Box 1053, Blindern, N-0316 Oslo, Norway (nilshr@ikaros.uio.no).

converges.) Other common models include Underwood's model,

$$v = v_{\max} e^{(-\rho/\rho_{\max})} \tag{0.4}$$

(which has been used as a model for low densities); Greenshield's model,

$$v = v_{\max}\left(1 - \left(\frac{\rho}{\rho_{\max}}\right)^n\right); \tag{0.5}$$

and the California model,

$$v = v_0\left(\frac{1}{\rho} - \frac{1}{\rho_{\max}}\right), \tag{0.6}$$

which, as the name suggests, fits best for heavy traffic. See [20, p. 296ff] and [7, p. 69ff] for an extensive comparison of these and other models and their experimental support. The resulting conservation law can be written as

$$\rho_t + f(\rho)_x = 0, \tag{0.7}$$

where

$$f(\rho) = \rho v(\rho). \tag{0.8}$$

Given initial data of the form $\rho(x,0) = \rho_0(x)$, (0.7) can be analyzed using the comprehensive theory available for conservation laws [24], [23], [17].

We will use a flux function f, which is concave with a unique maximum, and $f(0) = f(\rho_{\max}) = 0$, which is commonly assumed in traffic analysis ("the fundamental diagram of road traffic"); cf. [20], [7, p. 55], [16, p. 97], [9, p. 290], which excludes models (0.4) and (0.6). A modified analysis, however, would also cover these models. Before we discuss various extensions and alternative approaches to the analysis presented here, we will give a brief presentation of the model and our results.

Consider any finite, connected directed graph. Let the edges model the roads with traffic in the given direction. The vertices correspond to junctions. In addition, we may attach to the graph roads extending to infinity and connected to the graph at only one end. We call this system a *network*. The graph does not have to be planar; a nonplanar graph corresponds to the occurrence of bridges or tunnels. On each individual road we assume that we have heavy traffic in one direction only, so that the density satisfies a conservation law (0.7) for a certain flux function f. The roads connecting two junctions have finite length. If there is not much interaction between the traffic in the two directions on a two-way road or between the lanes in a multilane road, we can approximate this by two or several unidirectional roads connecting the same two endpoints (junctions). See Fig. 1.

A direct analysis of two-way traffic is complicated (see Bick and Newell [2]) since this leads to a system of conservation laws. Even in the simplest cases, one obtains a mixed hyperbolic/elliptic system for which even the solution of the Riemann problem is difficult; see, e.g., [10], [13].

One has to specify how the cars distribute at the junctions to obtain a well-defined model. We call this condition an *entropy condition* in accordance with the terminology used for conservation laws, since this condition gives a unique solution at least for Riemann initial data. In this paper we will discuss an entropy condition that comes from maximizing the flux at each intersection. More precisely, given a concave

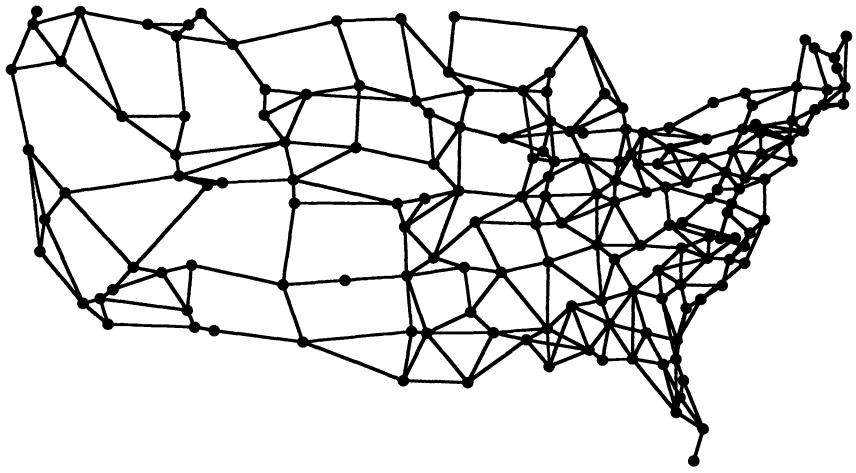


FIG. 1. *A traffic network.*

function g, we maximize the flow at the junction J measured by

$$\sum_{\text{roads } j \text{ at } J} g\left(\frac{f(\rho_j)}{f_{\max}}\right), \tag{0.9}$$

where $f_{\max}$ denotes the maximum of f.

We will now discuss why this is a reasonable entropy condition for this type of model. Consider a car going from a location A to a location B. When the car arrives at a junction, the driver will usually know which road to take and will make the appropriate choices at each junction. However, the continuum model for traffic that we use *does not* allow for individual tracking of cars, but it does describe the dynamics of the macroscopic density of the cars. What we have in mind is the following situation: if the driver arrives at a junction with the intention of turning, e.g., right, but sees that this road is (almost) blocked, she or he would probably take another road at this intersection (the principle of least resistance), and at a later intersection, after a small detour, get back on the right track. This also corresponds to the commonly observed fact the many people prefer to drive a longer distance at a higher speed than risk a severe bumper-to-bumper traffic jam. We model this by trying to maximize the flow through the junction locally for each intersection, which can be considered an important task for city planners who cannot consider individual cars *per se*. The timing of traffic lights, although not explicitly built into our model, also has the intention of maximizing the flow. Hence the model is probably most suitable for a heavily trafficked city center with many intersections. Our entropy condition will thus, in an average sense, maximize the flow at each junction as measured by the function g above. We imagine that the function g should be determined from observations and possibly depend on the intersection (although we use the same g for simplicity). We have not considered the question of how to measure g. Although our choice of entropy condition gives a unique solution to the problem with Riemann initial data, we have not proved uniqueness for the general Cauchy problem.

The solution of a conservation law develops singularities in finite time even for smooth initial data, thereby making it necessary to consider weak solutions of the partial differential equation.

Singularities in the sense of jump discontinuities moving with (finite) speed are called shocks. Specializing to Riemann initial data, i.e., two constant states separated by a single jump (see (2.1)), the solution consists of combinations of two elementary waves: shock waves and rarefaction waves. (This terminology comes from gas dynamics.) Rarefaction waves are continuous solutions.

One carefully has to distinguish between the concept of wave speed (e.g., the speed of a shock) and the speed of an individual car. In the context of traffic flow this can be exemplified as follows.

When a car on a road with heavy traffic has to reduce its speed due to, e.g., a traffic light turning red, we will have a shock wave, i.e., an abrupt change in density propagating in the opposite direction of the individual cars. Conversely, as the red light turns green, we will have rarefaction waves extending in both directions as the cars in front accelerate and spread out and the density of cars decreases continuously.

The method we use to solve the initial-boundary value problem for the resulting system of conservation laws is based on the technique introduced by Dafermos [4], where one replaces the flux function by a polygon, i.e., a continuous piecewise linear approximation. In addition, one approximates the initial value function by a step function, thereby obtaining (multiple) Riemann initial data. Since the flux function is piecewise linear, one obtains no rarefaction waves in the solution, since these are approximated by small shocks. These shocks are then *tracked*. Dafermos's method was proved to converge in [11], [12].

Thus, we first prove that the Riemann problems have a unique weak solution. The general Cauchy problem is approximated by finitely many Riemann problems, and using estimates from [11] we can prove existence of a solution of the general Cauchy problem:

The general framework presented in this paper allows for many extensions. We have assumed that the flux functions are identical for all the roads in the network, and, similarly, for the entropy function g. One could allow for individual flux functions for each road, thereby studying, for example, how the quality of individual roads as measured by the flux function influence traffic on the network. Furthermore, one could study the situation where the flux function also depends on position on the road, i.e., $f = f(x, \rho)$ (see, e.g., [19], [6], [21]), which could be used as a model of bottlenecks. In addition, traffic lights could be built into the model; see [19], [24].

Recently the study of stochastic conservation laws began; see [14], [15]. In this context, this would correspond to the situation where, e.g., the initial density is not known exactly or the exact form of f is uncertain. The analysis of such a stochastic model would be an interesting extension of the analysis in this paper.

A very interesting question to discuss for this type of problem is the possible occurrence of Braess's paradox [3], [1], [5].

For extensive discussions of the advantages and disadvantages of the hydrodynamical approach to traffic flow, we refer to [20], [7], [16], [9], which also contain alternative models for traffic flow. To the best of our knowledge, the hydrodynamical approach has not been analyzed for a network of roads prior to this analysis. Standard traffic models on networks treat the dynamics in a much simpler way than this model without using differential equations. These models then allow for discussing other questions resulting from, e.g., the tracking of individual cars. On the other hand, our model takes a more global point of view, assuming, in the sense discussed above, that the drivers try to maximize the flow locally at each intersection. From this and the continuum description, the model gives the total flow pattern. In this paper we dis-

cuss fundamental mathematical results of this model, provide a numerical technique, and give some examples on simple networks. We plan to treat a detailed, real-world example in a separate paper.

1. The Riemann problem. Consider a network of unidirectional roads which are connected at intersections. The precise definition of a network can be written as follows.

DEFINITION 1.1. *By a network we mean a finite, connected directed graph, where, in addition, we may attach a finite number of directed curves extending to infinity.*

Let each edge or curve correspond to a road and assume that the traffic is in the direction given by the direction in the network. The vertices correspond to junctions or intersections. A road may be connected to other roads at both ends or at one end only; in the latter case the unconnected end is assumed to extend to infinity. We assume that we have have N roads which are connected at M junctions. See Fig. 2.

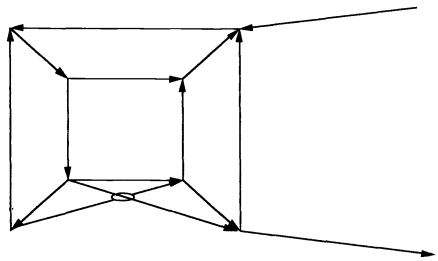


FIG. 2. *A network.*

Let $\rho_i = \rho_i(x,t)$ denote the density of cars at road i with $x \in [a_i, b_i]$, where $a_i = -\infty$ or $b_i = \infty$ if the road extends to infinity. Assume that the traffic goes in the direction of a_i to b_i. We will use the notation

(1.1)
$$\rho(x_1,\dots,x_N,t) = (\rho_1(x_1,t),\dots,\rho_N(x_N,t)), \quad t \geq 0, \quad x_i \in [a_i,b_i], \quad i = 1,\dots,N.$$

Away from intersections, conservation of cars yields [24], [17],

$$\frac{\partial \rho_i}{\partial t} + \frac{\partial f(\rho_i)}{\partial x} = 0, \quad i = 1,\dots,N, \tag{1.2}$$

which should be interpreted in the distributional sense. We have given an initial density at each road, viz.,

$$\rho_i(x,0) = \rho_{i,0}(x), \quad x \in [a_i,b_i], \quad i = 1,\dots,N. \tag{1.3}$$

The roads will be coupled in terms of boundary conditions at the intersections. The weak formulation of the system of conservation laws will provide us with the right boundary conditions.

Consider a junction J and assume, by relabeling if necessary, that roads $1,\dots,n$ enter J and roads $n+1,\dots,n+m$ leave J. (Whenever we study a fixed intersection J in the following, we will assume this labeling.) Write

$$c_i = \begin{cases} b_i & \text{for } i = 1,\dots,n, \\ a_i & \text{for } i = n+1,\dots,n+m. \end{cases} \tag{1.4}$$

Let $\phi = \{\phi_i\}_1^N$ be smooth test functions with ϕ_i defined on $[a_i, b_i]$ having compact support in $[a_i, b_i]$, $i = 1, \dots, n+m$, that also are C^1 smooth across junctions; i.e.,

$$\phi_i(c_i, t) = \phi_j(c_j, t) \quad \text{and} \quad \frac{\partial \phi_i}{\partial x}(c_i, t) = \frac{\partial \phi_j}{\partial x}(c_j, t) \tag{1.5}$$

for all $1 \le i, j \le n+m$, and similarly for other junctions. A weak solution of (1.2) is a set of functions $\rho_i(x,t)$, which satisfies

$$\sum_{i=1}^{N} \Big(\int_0^\infty \int_{a_i}^{b_i} \left[\rho_i \frac{\partial \phi_i}{\partial t} + f(\rho_i) \frac{\partial \phi_i}{\partial x} \right] dx dt + \int_{a_i}^{b_i} \rho_{i,0}(x) \phi_i(x,0)\, dx \Big) = 0 \tag{1.6}$$

for all ϕ satisfying (1.5).

By performing an analysis similar to the derivation of the Rankine–Hugoniot jump relation for shocks (see, e.g., [23, p. 246ff]), one easily obtains, as a consequence of the weak formulation, that for each intersection J,

$$\sum_{i=1}^{n} f\big(\rho_i(b_i, \cdot)\big) = \sum_{i=n+1}^{n+m} f\big(\rho_i(a_i, \cdot)\big) \tag{1.7}$$

for $t > 0$; i.e., all cars that enter an intersection must emerge at some other road leaving the intersection. (Consider test functions with support near the junction J and perform an integration by parts.) We call (1.7) the Rankine–Hugoniot condition for intersections. (Equation (1.7) resembles Kirchhoff's law in electromagnetism, but we will not use that terminology here.) Alternatively, we could consider the system (1.2) of N separate equations and impose boundary conditions (1.7) at each junction.

We will make certain assumptions on the function f that are reasonable in the context of traffic flow (cf. the discussion in the introduction). Normalize ρ so that $\rho_{\max} = 1$. We essentially assume that f has a unique maximum, viz.,

$$(\mathcal{F}) \qquad f(0) = f(1) = 0, \quad \exists \sigma \in \langle 0, 1 \rangle : f'(\sigma) = 0, \quad (\rho - \sigma) f'(\rho) < 0, \quad \rho \neq \sigma.$$

These restrictions hold for flux functions commonly used in the analysis of traffic flow [20], [9]. In the remaining part of this section we study the Riemann problem for a single junction J, i.e., the case with $M = 1$.

DEFINITION 1.2. *By the weak solution of the Riemann problem for the junction J, we mean the weak solution of the initial value problem* (1.2), (1.3), (1.7) *for the network consisting of the single junction J with n incoming roads and m outgoing roads, all extending to infinity. The initial data are given by*

$$\rho_{i,0}(x) = \rho_{i,0}, \quad x \in [a_i, b_i], \quad i = 1, \dots, N, \tag{1.8}$$

where $\rho_{i,0}$, $i = 1, \dots, N$ are constants and $N = n + m$. See Fig. 3.

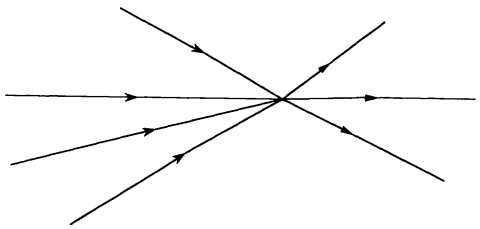


FIG. 3. *The Riemann problem.*

Before we introduce the notion of an entropy condition for the Riemann problem, we will discuss some immediate properties of any weak solution. Consider Riemann initial data (1.8) for a single junction. Irrespective of whether the initial data already satisfies the Rankine–Hugoniot condition at J, the solution must. But it is not sufficient to impose (1.7) to determine a unique solution since we have N unknowns (the possible states at the junction) and only one equation, the Rankine–Hugoniot condition. To obtain uniqueness, one has to impose a proper entropy condition. Let

$$\rho_J(t) = (\bar{\rho}_1, \dots, \bar{\rho}_N) = \rho(c_1, \dots, c_N, t) \tag{1.9}$$

denote the solution at the junction; thus, e.g., $\rho_J(0) = (\rho_{1,0}, \dots, \rho_{N,0})$. Once ρ_J is determined, we will solve the Riemann problem in the usual way for each of the roads with $\bar{\rho}_i$ as the right state for incoming roads ($i \le n$) and left state for outgoing roads ($i \ge n+1$). The solution will then consist of waves (either shock waves or rarefaction waves) emerging from the intersection J. For roads with incoming traffic (i.e., roads $i = 1, \dots, n$), these waves must have negative speed and, similarly, for the roads with outgoing traffic (i.e., roads $i = n+1, \dots, N$), the waves must have positive speed. Thus we see that for the Riemann problem, ρ_J will in fact be independent of t for $t > 0$. This imposes certain restrictions on the possible values of the solution close to the junction. For any $\rho \in [0,1]$, $\rho \neq \sigma$, we define $\tau(\rho)$ as the unique number with $\tau(\rho) \neq \rho$ such that $f(\tau(\rho)) = f(\rho)$. If $\rho \gtrless \sigma$ then $\tau(\rho) \lesseqgtr \rho$. Furthermore, $\tau(\sigma) = \sigma$. To be specific, consider an incoming road, i.e., $i \in \{1, \dots, n\}$. If $\rho_{i,0} > \sigma$ then $\bar{\rho}_i \in [\sigma, 1]$, while if $\rho_{i,0} < \sigma$ then $\bar{\rho}_i \in \{\rho_{i,0}\} \cup [\tau(\rho_{i,0}), 1]$. For outgoing roads, i.e., $i \in \{n+1, \dots, N\}$, we similarly have that

$$\bar{\rho}_i \in \begin{cases} [0, \sigma] & \text{for } \rho_{i,0} < \sigma, \\ [0, \tau(\rho_{i,0})] \cup \{\rho_{i,o}\} & \text{for } \rho_{i,0} > \sigma. \end{cases} \tag{1.10}$$

Let

$$A = I \times O \subseteq \tilde{A} = \prod_{i=1}^{n} [\sigma, 1] \times \prod_{i=n+1}^{n+m} [0, \sigma] \subset \mathbb{R}^{n+m}, \tag{1.11}$$

where

$$I = \prod_{i=1}^{n} [\tau(\rho_{i,0}), 1], \quad O = \prod_{i=n+1}^{n+m} [0, \tau(\rho_{i,0})]. \tag{1.12}$$

We would like to have $\rho_J \in A$, but because of the possibility of $\bar{\rho}_i = \rho_{i,0}$, if $\rho_{i,0} < \sigma$, and $i \in \{1 \dots, n\}$, and similarly for $i > n$, this is not guaranteed. However, we will see that we can modify the weak solution (on a set of measure zero) to obtain this. To see this, assume that we have a weak solution of the Riemann problem satisfying (1.6) as well as the Rankine–Hugoniot condition (1.7) and, to be specific, let us assume that $\bar{\rho}_1 = \rho_{1,0} < \sigma$. Then, the function

$$\hat{\rho}(x_1, \dots, x_N, t) = (\hat{\rho}_1(x_1, t), \dots, \hat{\rho}_N(x_N, t)) \tag{1.13}$$

with $\hat{\rho}_i = \rho_i$ for $i = 2, \dots, N$, and

$$\hat{\rho}_1(x_1, t) = \begin{cases} \tau(\rho_{1,0}) & \text{for } x_1 = 0, \\ \rho_1(x_1, t) & \text{for } x_1 < 0 \end{cases} \tag{1.14}$$

will still be a weak solution as $\hat{\rho} = \rho$ in $L^1(\prod_{i=1}^{N} [a_i, b_i] \times [0, \infty))$; it also satisfies the Rankine–Hugoniot condition. Here we have introduced *de facto* a stationary shock,

which we call a *virtual shock*, that will not change the solution. In this way we may (and we will from now on) assume that the solution at the junction is in A, i.e., $\rho_J \in A$.

This brings us to the point of introducing the entropy condition. The motivation for our choice was discussed in the introduction, and we will concentrate here on its mathematical formulation and consequences. The aim of this condition is to single out a unique point ρ_J in A such that the Rankine–Hugoniot condition (1.7) is satisfied. Having determined the value of the solution at the junction, we then determine the solution inside each of the intervals $[a_i, b_i]$, $i = 1, \dots, N$, in the standard way; see [23, p. 301ff]. (See the last section of this paper for an explicit example.)

Let the flux function f satisfy $(\mathcal{F})$, and let g be some differentiable strictly concave function of a single variable defined on $\mathbb{R}$. Write $\gamma_i = \gamma(\bar{\rho}_i) = f(\bar{\rho}_i)/f(\sigma)$ for $i = 1, \dots, N$. We define the entropy of the junction J as

$$E_J = \sum_{i=1}^{N} g(\gamma_i). \tag{1.15}$$

Introduce the Hugoniot locus (i.e., the points where the Rankine–Hugoniot condition is satisfied) as

$$H_J = \left\{ \rho_J = (\bar{\rho}_1, \dots, \bar{\rho}_N) \in \mathbb{R}^N \;\middle|\; \sum_{i=1}^{n} f(\bar{\rho}_i) = \sum_{i=n+1}^{N} f(\bar{\rho}_i) \right\}. \tag{1.16}$$

The entropy condition is written as follows:

$(\mathcal{E})$ *Find a* $\bar{\rho}_J = (\bar{\rho}_1, \dots, \bar{\rho}_N) \in H_J \cap A$ *which maximizes* E_J.

With the above assumptions on g we have that E_J is a strictly concave function defined on a convex set, and hence has a unique maximum. This maximum will be found either as an interior point where $\nabla E_J = 0$, or at the boundary of the domain. Let us first study stationary points. Using the function γ, we can write the Rankine–Hugoniot condition as

$$\sum_{i=1}^{n} \gamma_i = \sum_{i=n+1}^{N} \gamma_i \tag{1.17}$$

or

$$\gamma_N = \sum_{i=1}^{n} \gamma_i - \sum_{i=n+1}^{N-1} \gamma_i. \tag{1.18}$$

Observe that when $\rho_J \in A$ there is a one-to-one correspondence between γ_i and $\bar{\rho}_i$ which we will use whenever convenient in the rest of the paper. We may rewrite the entropy condition as

$$E_J = \sum_{i=1}^{N-1} g(\gamma_i) + g\left(\sum_{i=1}^{n} \gamma_i - \sum_{i=n+1}^{N-1} \gamma_i \right). \tag{1.19}$$

Thus

$$\frac{\partial E_J}{\partial \gamma_i} = g'(\gamma_i) + \frac{\partial \gamma_N}{\partial \gamma_i} g'(\gamma_N), \tag{1.20}$$

where

$$\frac{\partial \gamma_N}{\partial \gamma_i} = \begin{cases} +1 & \text{for } i = 1, \dots, n, \\ -1 & \text{for } i = n+1, \dots, N-1. \end{cases} \tag{1.21}$$

Therefore, $\nabla E_J = 0$ implies that $g'(\gamma_i) = -g'(\gamma_N)$ for $i = 1, \dots, n$, and $g'(\gamma_i) = g'(\gamma_N)$ for $i = n+1, \dots, N-1$. Thus

$$\gamma_i = \begin{cases} \frac{m}{n}\gamma_N & \text{for } i = 1, \dots, n, \\ \gamma_N & \text{for } i = n+1, \dots, N-1, \end{cases} \tag{1.22}$$

where γ_N is the solution of the equation

$$g'\left(\frac{m}{n}\gamma_N\right) = -g'(\gamma_N), \tag{1.23}$$

provided this equation has a solution $(m/n)\gamma_N$, $\gamma_N \in [0,1]$. (A unique solution of (1.23) will exist in $[0,1]$ if, e.g., $g'(0) = -g'(1)$.) The condition that $\rho_J \in A$ transforms into a condition $\gamma_J = (\gamma_1, \dots, \gamma_N) \in \prod_{i=1}^N [0, \kappa_i(\rho_{i,0})] \subseteq \prod_{i=1}^N [0,1]$. Here $\kappa(\rho_i)$ is defined as

$$\kappa_i(\rho) = \begin{cases} 1 & \text{if } i \le n \text{ and } \rho \ge \sigma, \text{ or } i > n \text{ and } \rho \le \sigma, \\ \gamma(\rho) & \text{otherwise.} \end{cases} \tag{1.24}$$

If the stationary point we have computed above is not in $\prod_{i=1}^N [0, \kappa_i(\rho_{i,0})]$, one or more of the coordinates will have to be found on the boundary. We omit the details.

As we said above, once we have determined the (unique) value of the solution at the junction, we construct the solution on the individual roads. We have now proved the following theorem.

THEOREM 1.1. *Assume that f satisfies $(\mathcal{F})$, and let g be a strictly concave differentiable function. Then the Riemann problem for the single junction J with entropy condition $(\mathcal{E})$ has a unique solution. Let ρ_J denote the entropy solution at the junction. Then the solution is written as follows: First consider incoming roads $(i \in \{1, \dots, n\})$. For $\bar{\rho}_i > \rho_{0,i}$, we have*

$$\rho_i(x_i, t) = \begin{cases} \rho_{i,0} & \text{for } x < st, \\ \bar{\rho}_i & \text{for } st < x < 0, \end{cases} \tag{1.25}$$

while if $\bar{\rho}_i < \rho_{0,i}$,

$$\rho_i(x_i, t) = \begin{cases} \rho_{i,0} & \text{for } x < f'(\rho_{i,0})t, \\ (f')^{-1}(\frac{x}{t}) & \text{for } f'(\rho_{i,0})t < x < tf'(\bar{\rho}_i), \\ \bar{\rho}_i & \text{for } tf'(\bar{\rho}_i) < x < 0. \end{cases} \tag{1.26}$$

For outgoing roads $(i \in \{n+1, \dots, N\})$, we similarly find that if $\bar{\rho}_i < \rho_{0,i}$,

$$\rho_i(x_i, t) = \begin{cases} \bar{\rho}_{i,0} & \text{for } 0 < x < st, \\ \rho_{i,0} & \text{for } st < x, \end{cases} \tag{1.27}$$

while if $\bar{\rho}_i > \rho_{0,i}$, we have

$$\rho_i(x_i, t) = \begin{cases} \bar{\rho}_i & \text{for } 0 < x < f'(\bar{\rho}_i)t, \\ (f')^{-1}(\frac{x}{t}) & \text{for } f'(\bar{\rho}_i)t < x < tf'(\rho_{i,0}), \\ \rho_{i,0} & \text{for } tf'(\rho_{i,0}) < x. \end{cases} \tag{1.28}$$

Here $s = [f]/[\rho] = (f(\rho_{i,0}) - f(\bar{\rho}_i))/(\rho_{i,0} - \bar{\rho}_i)$ *is the (standard) Rankine–Hugoniot shock speed.*

2. Construction of approximate solutions. We will now turn our attention to the general Cauchy problem with the entropy condition $(\mathcal{E})$ imposed on each intersection. Based on Dafermos's method [4], [11], [12], we will construct approximate solutions to this problem, and in the next section we will show that this construction yields a convergent subsequence.

First, we briefly review the solution of the Riemann problem for the scalar conservation law if the flux function f is piecewise linear and concave. Consider the Riemann problem

$$\begin{aligned} &u_t + f(u)_x = 0, \\ &u(x,0) = \begin{cases} u_\ell & \text{for } x < 0, \\ u_r & \text{for } x \geq 0, \end{cases} \end{aligned} \tag{2.1}$$

where u is a scalar function. If $u_\ell \leq u_r$ then the solution is simply

$$u(x,t) = \begin{cases} u_\ell & \text{for } x < st, \\ u_r & \text{for } x > st, \end{cases} \tag{2.2}$$

where $s = [f]/[u] = (f(u_\ell) - f(u_r))/(u_\ell - u_r)$. If $u_\ell > u_r$, we have the following solution: let $u_0 = u_\ell$, $u_k = u_r$, and

$$u_0 < u_1 < \cdots < u_k, \quad f_i = f(u_i) \in \mathbb{R}, \quad i = 0, \ldots, k, \tag{2.3}$$

such that f is linear on each $[u_i, u_{i+1}]$, viz.,

$$u \in [u_i, u_{i+1}] \Rightarrow f(u) = \frac{f_{i+1} - f_i}{u_{i+1} - u_i}(u - u_i) + f_i, \quad i = 0, \ldots, k-1. \tag{2.4}$$

Then we have the following theorem.

THEOREM 2.1. *Consider the Riemann problem* (2.1), *where* f *is piecewise linear and concave. Then the solution is given by*

$$u(x,t) = \begin{cases} u_\ell & \text{for } x < st, \\ u_r & \text{for } x > st \end{cases} \tag{2.5}$$

if $u_\ell \leq u_r$, *where* $s = [f]/[u] = (f(u_\ell) - f(u_r))/(u_\ell - u_r)$ *and*

$$u(x,t) = \begin{cases} u_\ell & \text{for } x < s_0 t, \\ u_i & \text{for } s_{i-1} t < x < s_i t, \quad i = 1, \ldots, k-1, \\ u_r & \text{for } x > s_{k-1} t, \end{cases} \tag{2.6}$$

where

$$s_i = \frac{f_{i+1} - f_i}{u_{i+1} - u_i}, \quad i = 0, \ldots, k-1 \tag{2.7}$$

if $u_\ell > u_r$.

Observe that in this case $u(x,t)$ is a piecewise constant function in x/t. For a scalar conservation law on the line, Dafermos's scheme consists of approximating the initial function by a step function, thereby generating a series of Riemann problems. The solution of these problems will then define a set of discontinuities which propagate linearly in (x,t) space. At some $t_1 > 0$, two of these discontinuities will collide, and one then solves the Riemann problem defined by the values to the right and left of the

collision. This gives another set of discontinuities, and the process can be continued up to the next collision time. In [11] it is shown that this is a well-defined process and that the solutions generated converge as the approximation of the flux function f and the initial value function u_0 converge.

We will use this strategy to construct an approximation of the solution of (1.2), (1.3). First, we make a polygonal approximation of f. Let k be some positive even integer and divide the interval $[0, \sigma]$ into $k/2+1$ intervals of length $2\sigma/k$. The interval $\langle \sigma, 1]$ is then divided at points $\{\tilde{\rho}_j\}_{j=k/2+1}^{k}$, chosen such that

$$f\left(2\frac{j}{k}\sigma\right) = f(\tilde{\rho}_{k-j}) \tag{2.8}$$

for $j = 0, \dots, k/2$, i.e., $\tilde{\rho}_j = \tau((2\sigma/k)j)$. For $j \leq k/2$ we define $\tilde{\rho}_j = 2(j/k)\sigma$. We define $f^k(\rho)$ as the piecewise linear continuous function

$$f^k(\rho) = f(\tilde{\rho}_j) + \frac{f(\tilde{\rho}_j) - f(\tilde{\rho}_{j+1})}{\tilde{\rho}_j - \tilde{\rho}_{j+1}}(\rho - \tilde{\rho}_j) \quad \text{for } \rho \in [\tilde{\rho}_j, \tilde{\rho}_{j+1}]. \tag{2.9}$$

Having approximated the flux function f, we now approximate the initial data $\rho_{i,0}$ with a step function $\rho_{i,0}^k(x)$ taking values in the set $\{\tilde{\rho}_j\}_{j=0}^{k}$. The solution of the Riemann problem (2.1) away from the intersections will take values in $\{\tilde{\rho}_j\}$ (cf. Theorem 2.1), but the solution of the Riemann problem (1.8) at the junctions may take us outside $\{\tilde{\rho}_j\}$. However, we approximate the exact solution by choosing the closest density in this set; i.e., assume that on one particular road the exact solution of the Riemann problem adjacent to the junction is given by $\rho \in \langle \tilde{\rho}_j, \tilde{\rho}_{j+1} \rangle$. Then we let the approximate solution be $\tilde{\rho}_j$ if $|\tilde{\rho}_j - \rho| < |\tilde{\rho}_{j+1} - \rho|$ and $\tilde{\rho}_{j+1}$ otherwise. By doing this, we introduce an error in conservation, and the approximate solution is no longer an exact weak solution of an approximate problem. But this error is of size $O(1/k)$, and thus vanishes as $k \to \infty$.

In the remaining part of this paper we let N denote the total number of roads in the network, and we also denote the densities on all roads by the vector $\rho = (\rho_1 \dots, \rho_N)$, calling attention to the difference between the break point $\tilde{\rho}_j$ and the function $\rho_j(x,t)$. The approximate solution is then denoted by $\rho^k = (\rho_1^k, \dots, \rho_N^k)$. For the Riemann problems defined by $f = f^k$ and the discontinuities of $\rho_{i,0}^k$, we can find the solution using Theorem 2.1. The approximate solution $\rho^k = (\rho_1^k, \dots, \rho_N^k)$ defines a set of discontinuities moving in the intervals $[a_i, b_i]$ for $i = 1, \dots, N$. Clearly, ρ^k can be defined at least until the first collision of discontinuities. This collision defines a new Riemann problem whose solution creates new discontinuities which can be propagated until the next collision and so on. Obviously, this can be repeated an arbitrary number of times. Below we will show that ρ^k can indeed be defined up to any time, but first we must show some lemmas.

We now have that $\rho_i^k(x,t)$, $i = 1, \dots, N$, will be a step function in x and thus will define a number of constant states $\rho_{i,j}^k$ for $j = 0, \dots, n_i$ such that $\rho_{i,0}^k = \rho_i^k(a_i, \cdot)$ and so on. We will also label the times (in increasing order) when some discontinuities collide, either with each other or with an intersection, by t_ℓ for $\ell = 1, 2, \dots$.

LEMMA 2.1. *Assume that we have an entropy state ρ_J (and the corresponding $\gamma_J = (\gamma_1, \dots, \gamma_N)$), which is a solution of the Riemann problem for the junction J. Consider a shock colliding with J from road r. Let the new entropy solution after the*

collision be denoted by $\hat{\rho}_J$ *with corresponding* $\hat{\gamma}_J = (\hat{\gamma}_1, \dots, \hat{\gamma}_N)$. *If* $r \le n$ *then*

$$\begin{aligned} &\hat{\gamma}_i \ge \gamma_i \quad \text{for } i = 1, \dots, n,\ i \ne r, \\ &\hat{\gamma}_i \le \gamma_i \quad \text{for } i = n+1, \dots, n+m, \end{aligned} \tag{2.10}$$

and if $r > n$ *then*

$$\begin{aligned} &\hat{\gamma}_i \le \gamma_i \quad \text{for } i = 1, \dots, n, \\ &\hat{\gamma}_i \ge \gamma_i \quad \text{for } i = n+1, \dots, n+m,\ i \ne r. \end{aligned} \tag{2.11}$$

Proof. We will show the lemma in the case $r \le n$; the proof in the other case is identical. Since the colliding discontinuity has positive speed, the density to the left of the discontinuity $\hat{\rho}$ must be in the set $\{\rho | \rho < \sigma \text{ and } \gamma(\rho) < \gamma_r\}$. The possible densities adjacent to the junction on road r after the collision are from the set $\{\rho | \rho > \sigma \text{ and } \gamma(\rho) \le \gamma(\hat{\rho})\}$. In particular, $\hat{\gamma}_r < \gamma_r$. Now (2.10) will follow if

$$\frac{\partial E_J}{\partial \gamma_i}\bigg|_{(\gamma_1, \dots, \hat{\gamma}_r, \dots, \gamma_{m+n-1})} \begin{cases} > 0 & \text{for } i = 1, \dots, n,\ i \ne r, \\ < 0 & \text{for } i = n+1, \dots, n+m. \end{cases} \tag{2.12}$$

Let $\hat{\gamma}_{n+m} = \sum_{i \le n, i \ne r} \gamma_i - \sum_{i > n} \gamma_i + \hat{\gamma}_r$. Since $\hat{\gamma}_r < \gamma_r$, $\hat{\gamma}_{n+m} < \gamma_{n+m}$, and since g' is a decreasing function, $g'(\hat{\gamma}_{n+m}) > g'(\gamma_{n+m})$. Thus, from (1.21) we get, for $i \le n$, that

$$0 = \frac{\partial E_J}{\partial \gamma_i}\bigg|_{\bar{\gamma}} = g'(\gamma_i) + g'(\gamma_{n+m}) < g'(\gamma_i) + g'(\hat{\gamma}_{n+m}) = \frac{\partial E_J}{\partial \gamma_i}\bigg|_{(\gamma_1, \dots, \hat{\gamma}_r, \dots, \gamma_{m+n-1})} \tag{2.13}$$

Similarly, for $i > n$,

$$\frac{\partial E_J}{\partial \gamma_i}\bigg|_{(\gamma_1, \dots, \hat{\gamma}_r, \dots, \gamma_{m+n-1})} < 0, \tag{2.14}$$

and the lemma is proved. □

LEMMA 2.2. *Assume that we have an entropy state* ρ_J *with corresponding* $\gamma_J = (\gamma_1, \dots, \gamma_N)$, *which is a solution of the Riemann problem for the junction* J. *Consider a shock colliding with* J *from road* r. *Let the new entropy solution after the collision be denoted by* $\tilde{\rho}_J$ *with corresponding* $\tilde{\gamma}_J = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_N)$. *Then*

$$|\gamma_r - \tilde{\gamma}_r| = \sum_{i \ne r} |\gamma_i - \tilde{\gamma}_i|. \tag{2.15}$$

Proof. Without loss of generality we can assume that $r \le n$. By the previous lemma, $\tilde{\gamma}_i \ge \gamma_i$ for $i \le n$, $i \ne r$, and $\tilde{\gamma}_i \le \gamma_i$. Subtracting the Rankine–Hugoniot relations for γ_J and $\tilde{\gamma}_J$ we get

$$\gamma_r - \tilde{\gamma}_r = \sum_{i \ne r}^{n} (\tilde{\gamma}_i - \gamma_i) - \sum_{i=n+1}^{n+m} (\tilde{\gamma}_i - \gamma_i) = \sum_{i \ne r}^{n} (\tilde{\gamma}_i - \gamma_i) + \sum_{i=n+1}^{n+m} (\gamma_i - \tilde{\gamma}_i). \tag{2.16}$$

All terms in the last equation are positive and the lemma follows. □

Introduce

$$z = z(\rho) = (\gamma(\rho), \operatorname{sgn}(\rho - \sigma)) \tag{2.17}$$

with $\operatorname{sgn}(x) = -1$ for $x \le 0$ and 1 otherwise, and write, for simplicity,

$$z_{i,j}^k = z(\rho_{i,j}^k). \tag{2.18}$$

Furthermore, for $t_{\ell-1} < t < t_\ell$, we define the functional T^ℓ by

$$T^\ell = \sum_{\nu=1}^{N} T_\nu^\ell = \sum_{\nu=1}^{N} \sum_{i=1}^{n_\nu} |z_{\nu,i}^k - z_{\nu,i-1}^k|, \tag{2.19}$$

where $|(f,i)| = |f| + |i|$ and n_ν is the number of discontinuities on road ν. We now have the following lemma.

LEMMA 2.3.

$$T^{\ell+1} \le T^\ell. \tag{2.20}$$

Proof. It follows from Theorem 3.1 in [11] that T does not increase when two discontinuities collide. In the case where a discontinuity collides with a junction, Lemma 2.2 implies that T does not increase. This is true because the waves emitted from a junction in the approximated solution will always be smaller or equal to the correct waves described by Lemma 2.2. □

Now, since γ takes values in a finite set and T is positive, it follows that after some finite number of collisions L, T remains equal to some constant T^L. If T is constant for collisions between discontinuities, we can only have collisions between discontinuities separating (γ_ℓ, i_ℓ) and (γ_m, i_m), and (γ_m, i_m) and (γ_r, i_r). In this case γ_m must be between γ_ℓ and γ_r. In particular, this implies that both discontinuities have positive or negative speed. So, after t_L all roads must have discontinuities of either positive or negative speed only. Let us examine the situation on some particular road r after t_L. Assume that all discontinuities on r have positive speed. From the start of r, a single discontinuity can emerge because of a collision with the junction along some other road connected to r there. No discontinuity can emerge from the end of r, since this would lead to a collision between fronts which have speeds of different signs. Thus, eventually, all fronts on r will have collided with the right endpoint of r, and we are left with a single virtual shock. The situation is similar if all speeds on r are negative. In particular, front tracking on a traffic network is a well-defined construction.

3. The Cauchy problem. In this section we will show three lemmas which, by a standard diagonalization argument, imply that as $k \to \infty$ a subsequence of $z^k = (z_1^k, \dots, z_N^k)$ converges locally in $L_1(\prod_{i=1}^N [a_i, b_i] \times [0, \infty\rangle)$.

LEMMA 3.1.

$$\sup_{k,i,x,t} |z_i^k(x,t)| \le 2. \tag{3.1}$$

This is obvious and needs no proof.

Now we will define the total variation of z^k:

$$\text{T.V.}(z^k) = \sum_{i=1}^{N} \text{T.V.}_{x\in[a_i,b_i]} z_i^k(x,t). \tag{3.2}$$

Our second lemma states that the total variation is uniformly bounded in time.

LEMMA 3.2.

$$\text{T.V.}\big(z^k(x,t)\big) \le K \tag{3.3}$$

for some constant K independent of k and t.

Proof. By construction of the functional T^ℓ, (2.20), and Lemma 2.3, we have that

$$\begin{aligned} \text{T.V.}\big(z^k(x,t)\big) &\le T^\ell \quad \text{for } t \in \langle t_{\ell-1}, t_\ell], \\ &\le T^1 \le K, \end{aligned} \tag{3.4}$$

which proves the lemma. □

LEMMA 3.3. *Let $\tau_1 > \tau_2$ be such that $(\tau_1 - \tau_2)$ is sufficiently small. Then there is a constant K independent of k, i, τ_1, or τ_2 such that*

$$\int_{a_i}^{b_i} |z_i^k(x,\tau_1) - z_i^k(x,\tau_2)|dx \le K(\tau_1 - \tau_2). \tag{3.5}$$

Proof. Let r be a road with "density" $z^k(x,t)$ and endpoints a and b. Assume that r is connected at a through the junction J_a with roads $r_1, \dots, r_n$, and at b through the junction J_b with roads $r_{1+n}, \dots, r_{m+n}$. Let C be a number such that $C \ge \max|f'(\rho)|$ such that the speed of any discontinuity in $\bar{\rho}^k$ is bounded by C. Now assume that $(\tau_1 - \tau_2)$ is so small that $C(\tau_1 - \tau_2) \le \frac{1}{2}\min_{1\le i\le N}(b_i - a_i)$. Define $a' = a + C(\tau_1 - \tau_2)$ and $b' = b - C(\tau_1 - \tau_2)$. For $x < a'$, let $t_a(x) = \tau_1 - (x-a)/C$, and for $x > b'$ let $t_b(x) = \tau_1 - (b-x)/C$. For $a' \le x \le b'$ we have that $|z_i^k(x,\tau_1) - z_i^k(x,\tau_2)|$ is bounded by the spatial variation of $z_i^k(y,\tau_2)$, where $x - C(\tau_1 - \tau_2) < y < x + C(\tau_1 - \tau_2)$. For $x < a'$ or $x > b'$, $|z_i^k(x,\tau_1) - z_i^k(x,\tau_2)|$ is bounded by 4 by Lemma 3.1. Thus

$$\begin{aligned} \int_a^b |z^k(x,\tau_1) - z^k(x,\tau_2)|dx &\le 2(a'-a) + \int_{a'}^{b'} |z^k(x,\tau_1) - z^k(x,\tau_2)|dx + 2(b-b') \\ &\le \int_{a'}^{b'} \int_{x-C(\tau_1-\tau_2)}^{x+C(\tau_1-\tau_2)} \left|\frac{dz}{dy}\right| dydx + 8C(\tau_1 - \tau_2), \end{aligned} \tag{3.6}$$

where $|dz/dy|$ is a measure with total mass T.V.(z^k). By changing the order of integration we obtain for some constant $\tilde{C}$,

$$\int_a^b |z^k(x,\tau_1) - z^k(x,\tau_2)|dx \le \tilde{C}M(\tau_1 - \tau_2)\text{T.V.}(\rho^k), \tag{3.7}$$

where the right-hand side is bounded by Lemma 3.2. □

Now, by a standard technique (see, e.g., [23, p. 385ff]), one can show that Lemmas 3.1–3.3 imply that as $k \to \infty$, a subsequence of z^k converges in L_1, or, more precisely, we have the following theorem.

THEOREM 3.1. *Assume that f satisfies $(\mathcal{F})$. Let $z_0(x_1, \dots, x_N) = \big(z_{1,0}(x_1), \dots, z_{N,0}(x_N)\big)$ be such that* T.V.(z_0) *is bounded. Then, as $k \to \infty$, there exists a subsequence k_j and a function $z(x_1, \dots, x_N, t) = \big(z_1(x_1,t), \dots, z_N(x_N,t)\big)$ such that for any finite time T, z^{k_j}, as defined in §2, converges uniformly to z in L_1^{loc} for any $t < T$.*

Since the map (2.17), viz., $\Psi : \rho \mapsto z(\rho)$, is such that its inverse Ψ^{-1} is uniformly continuous, this implies that some subsequence of ρ^k also converges to $\rho = \Psi^{-1}(z)$.

Now we would like to show that the limit function ρ is indeed a weak solution to our problem (1.1), and to this end we have to show that the functional

$$W(\rho) = \sum_{i=1}^N \Big(\int_0^\infty \int_{a_i}^{b_i} \left[\rho_i \frac{\partial \phi_i}{\partial t} + f(\rho_i) \frac{\partial \phi_i}{\partial x} \right] dxdt + \int_{a_i}^{b_i} \rho_{i,0}(x)\phi_i(x,0)dx \Big) \tag{3.8}$$

vanishes for all ϕ_i in C_0^1 for $i = 1, \dots, N$. For simplicity we will now label our convergent subsequence ρ^k. Since ρ^k is almost a weak solution of (1.2) and, therefore,

satisfies $W(\rho^k) = O(N/k)$ with $f = f^k$, we can write

$$W(\rho) = \sum_{i=1}^{N} \Big(\int_0^\infty \int_{a_i}^{b_i} \left[(\rho_i - \rho_i^k) \frac{\partial \phi_i}{\partial t} + \big(f(\rho_i) - f^k(\rho_i^k)\big) \frac{\partial \phi_i}{\partial x} \right] dx\, dt + \int_{a_i}^{b_i} [\rho_{i,0}(x) - \rho_{i,0}^k(x)] \phi_i(x,0)\, dx \Big) + O\left(\frac{N}{k}\right) \tag{3.9}$$

for any k. Therefore,

$$|W(\rho)| \leq \begin{cases} \sum_{i=1}^N \|\frac{\partial \phi_i}{\partial t}\|_\infty \|\rho_i - \rho_i^k\|_{L_1} + \|\frac{\partial \phi_i}{\partial x}\|_\infty \|f(\rho_i) - f^k(\rho_i^k)\|_{L_1} \\ + \|\phi_i\|_\infty \|\rho_i(x,0) - \rho_i^k(x,0)\|_{L_1} \end{cases}$$

$$\leq \sum_{i=1}^{N} \left\| \frac{\partial \phi_i}{\partial t} \right\|_\infty \|\rho_i - \rho_i^k\|_{L_1} \tag{3.10a}$$

$$+ \left\| \frac{\partial \phi_i}{\partial x} \right\|_\infty \|f(\rho_i) - f^k(\rho_i)\|_{L_1} \tag{3.10b}$$

$$+ \left\| \frac{\partial \phi_i}{\partial x} \right\|_\infty \|f^k(\rho_i) - f^k(\rho_i^k)\|_{L_1} \tag{3.10c}$$

$$+ \|\phi_i\|_\infty \|\rho_i(x,0) - \rho_i^k(x,0)\|_{L_1} \tag{3.10d}$$

$$+ NO\left(\frac{1}{k}\right). \tag{3.10e}$$

By Lipshitz continuity of f and f^k, the terms (3.10b) and (3.10c) are small; the term (3.10a) is small since $\hat{\rho}$ is the L_1 limit of ρ^k, and the last term (3.10d) is small by the construction of $\rho^k(x_1, \ldots, x_N, 0)$. Therefore, for any $\epsilon > 0$, we may find a k such that $|W(\rho)| < \epsilon$, concluding that ρ is indeed a weak solution. Thus we have proven the following theorem.

THEOREM 3.2. *Consider the Cauchy problem* (1.2), (1.3), (1.7) *with flux function* f *satisfying condition* $(\mathcal{F})$. *Assume that the initial data* $\rho_0(x_1, \ldots, x_N) = (\rho_1(x_1), \ldots, \rho_N(x_N))$ *is such that* T.V. $(\Psi(\rho_0))$ *is bounded. Then the initial-boundary problem has a weak solution constructed from the solution of the Riemann problem using the entropy condition* $(\mathcal{E})$.

Example 1. Consider a simple junction consisting of two incoming roads and one outgoing road. Let the density on the two incoming roads be denoted by ρ_1 and ρ_2, and the density on the outgoing road be denoted by ρ_3. Similarly, we denote the fluxes γ_i. Let the flux function f be given by

$$f(\rho) = 4\rho(1 - \rho). \tag{3.11}$$

The choice of entropy function g will be motivated by practical insight in traffic flow, and this choice will of course determine the solution. As we shall see below, not all choices of g give reasonable results. Let us first use $g = f$. Assume that we initially have the entropy state $\gamma^{\mathcal{E}}$,

$$\gamma^{\mathcal{E}} = \left(\frac{1}{3}, \frac{1}{3}, \frac{2}{3}\right) \quad \text{or} \quad \bar{\rho} = \left(\frac{3+\sqrt{6}}{6}, \frac{3+\sqrt{6}}{6}, \frac{3-\sqrt{3}}{6}\right). \tag{3.12}$$

Now assume that a discontinuity with left density 0 collides with the junction from

road 1 at $t = 0$; i.e., for negative time, the density on road 1 is given by

$$\rho_1 = \begin{cases} 0 & \text{for } x < st, \\ \frac{3+\sqrt{6}}{6} & \text{for } x \geq st, \end{cases} \tag{3.13}$$

where $s = 2/(3+\sqrt{6})$ is the speed of the discontinuity given according to the Rankine–Hugoniot condition. Where there is no ambiguity, we will also denote the densities for $t > 0$ by ρ_i. After the collision, the density on road 1 can only be 0 away from the junction with a virtual shock at the endpoint connecting it to a density of 1. Therefore, the Rankine–Hugoniot condition now demands that $\gamma_2 = \gamma_3$ and $\rho_2 = 1 - \rho_3$. Since $g(\gamma)$ has a maximum for $\gamma = \frac{1}{2}$, the entropy condition $(\mathcal{E})$ gives $\gamma_2 = \gamma_3 = \frac{1}{2}$ or $\rho_2 = (2+\sqrt{2})/4$ and $\rho_3 = (2-\sqrt{2})/4$. Thus we see that the densities for positive time are given by

$$\begin{aligned} &\rho_1(x,t) = 0, \\ &\rho_2(x,t) = \begin{cases} \frac{3+\sqrt{6}}{6} & \text{for } x \leq -4\sqrt{\frac{2}{3}}t, \\ \frac{1}{2} - \frac{x}{8t} & \text{for } -4\sqrt{\frac{2}{3}}t < x \leq -2\sqrt{2}t, \\ \frac{2+\sqrt{2}}{4} & \text{for } -2\sqrt{2}t < x, \end{cases} \\ &\rho_3(x,t) = \begin{cases} \frac{2-\sqrt{2}}{4} & \text{for } x < s_3 t, \\ \frac{3-\sqrt{3}}{6} & \text{for } x \geq s_3 t. \end{cases} \end{aligned} \tag{3.14}$$

The speed s_3 is given by

$$s_3 = \frac{2}{3\sqrt{2} - 2\sqrt{3}}. \tag{3.15}$$

Thus, as expected, the densities adjacent to the junction on roads 2 and 3 are smaller after the collision than before, but we also note that this entropy function implies that the density adjacent to the junction becomes *smaller* than both the densities on roads 2 and 3. It would be reasonable to expect that once road 1 stops feeding cars into the junction, its effect would no longer be felt, and the solution for positive time would be the same as for the Riemann problem on a single road with initial data

$$\rho(x,0) = \begin{cases} \frac{3+\sqrt{6}}{6} & \text{for } x \leq 0, \\ \frac{3-\sqrt{3}}{6} & \text{for } x > 0. \end{cases} \tag{3.16}$$

If we choose $g(\gamma)$ as a function which has a maximum for $\gamma = 1$, then the above argument shows that after the collision we have $\rho_2 = \rho_3 = \frac{1}{2}$. Then the solution of the Riemann problem on each road is a rarefaction wave and these waves "fit" for $x/t = 0$. Thus, the solution is given by

$$\begin{aligned} \rho_2(x,t) &= \begin{cases} \frac{3+\sqrt{6}}{6} & \text{for } x \leq -4\sqrt{\frac{2}{3}}t, \\ \frac{1}{2} - \frac{x}{8t} & \text{for } -4\sqrt{\frac{2}{3}}t < x \leq 0, \end{cases} \\ \rho_3(x,t) &= \begin{cases} \frac{1}{2} - \frac{x}{8t} & \text{for } 0 \leq x < \frac{4}{\sqrt{3}}t, \\ \frac{3-\sqrt{3}}{6} & \text{for } \frac{4}{\sqrt{3}}t \leq x. \end{cases} \end{aligned} \tag{3.17}$$

We see that the junction no longer influences the densities, since this is also the solution of the Riemann problem (3.16). In practice, this choice of g would therefore be more

reasonable.

With this new choice of entropy function, the initial state $\gamma^{\mathcal{E}}$ does not satisfy the entropy condition; i.e., it is not a solution to the Riemann problem. If we assume that the fluxes are the same but the densities are smaller than $\sigma = \frac{1}{2}$ on the incoming roads, we have that $\gamma^{\mathcal{E}}$ is an entropy state.

Example 2. Now assume that we have a simple network consisting of a double fork with two incoming and two outgoing roads such that one of the outgoing roads doubles back into the network again; see Fig. 4. We will use the same f as in example 1 and assume that g is a strictly concave function with a maximum at $\gamma = 1$.

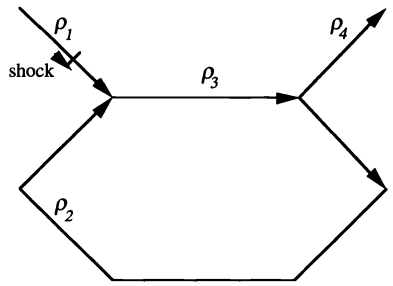


FIG. 4. *A traffic network.*

Denote the left junction by A and the right one by B. To simplify the discussion, we consider a situation with low density traffic with $\rho_{2,0} = \rho_{4,0} = (3 - \sqrt{6})/6$ and $\rho_{3,0} = (3 - \sqrt{3})/6$. The densities are adjusted so that both the Rankine–Hugoniot condition and the entropy condition are satisfied initially at B, and hence the solution at B will remain unchanged until one has interference from A. Now assume that we have a shock on road 1 with left value 0 that hits A at time $t = 0$, i.e., for $t < 0$,

$$\rho_1(x,t) = \begin{cases} \frac{3-\sqrt{6}}{6} & \text{for } st < x \leq 0, \\ 0 & \text{for } x \leq st \end{cases} \tag{3.18}$$

with speed $s = 2/(3 - \sqrt{6})$. Thus at the junction A we have that $\bar{\rho}_1(0) = \bar{\rho}_{1,0} = 0$, $\bar{\rho}_2(0) = \bar{\rho}_{2,0} = (3 - \sqrt{6})/6$, and $\bar{\rho}_3(0) = \bar{\rho}_{3,0} = (3 - \sqrt{3})/6$ at $t = 0$. We see from the velocity considerations prior to (1.10) that, using virtual shocks, the solution on road 1 can be written as

$$\rho_1(x,t) = \begin{cases} 1 & \text{for } x = 0, \\ 0 & \text{for } x < 0. \end{cases} \tag{3.19}$$

Thus $\gamma_1 = 0$ and $\gamma_2 = \gamma_3$. From the entropy condition we infer that $\gamma_2 = \gamma_3 = \frac{1}{3}$, giving $\bar{\rho}_2 = \tau(\bar{\rho}_{2,0})$ and $\bar{\rho}_3 = \bar{\rho}_{2,0}$, which implies that we have another virtual shock on road 2. Hence, near A for small t we have

$$\rho_2(x,t) = \begin{cases} \frac{3+\sqrt{6}}{6} & \text{for } x = b_2, \\ \frac{3-\sqrt{6}}{6} & \text{for } x < b_2 \end{cases} \tag{3.20}$$

(recall that b_2 denotes the end point of road 2), and

$$\rho_3(x,t) = \begin{cases} \frac{3-\sqrt{6}}{6} & \text{for } x < s_3 t, \\ \frac{3-\sqrt{3}}{6} & \text{for } x > s_3 t \end{cases} \tag{3.21}$$

with $s_3 = 1/(3\sqrt{3} - 3\sqrt{6})$.

At $t = t_0 = (b_3 - a_3)s_3$ the shock on road 3 will hit junction B. The "initial state" for $t = t_0$ at B will now be $\rho_2(a_2, t_0) = \rho_3(b_3, t_0) = \rho_4(a_4, t_0) = (3 - \sqrt{6})/6$. From the Rankine–Hugoniot condition we have that the solution has to satisfy $\gamma_3 = \gamma_2 + \gamma_4$, and, in addition, the entropy condition requires that $g(\gamma_3) + g(\gamma_2) + g(\gamma_4)$ be maximized with $\gamma_j \in [0, \kappa_j((3 - \sqrt{6})/6)] = [0, \frac{1}{3}]$ for $j = 2, 3, 4$. We infer that $\gamma_2 = \gamma_4$ because of symmetry, and hence we have to determine the maximum of $h(\gamma_2) = g(2\gamma_2) + 2g(\gamma_2)$ with $2\gamma_2, \gamma_2 \in [0, \frac{1}{3}]$. Since g is strictly concave with maximum at $\gamma = 1$, we see that the maximum is reached when $2\gamma_2 = \gamma_3 = \frac{1}{3}$, and hence $\gamma_2 = \gamma_4 = \frac{1}{6}$, which implies that $\bar{\rho}_2 = \bar{\rho}_4 = \frac{1}{2}(1 - \sqrt{5/6})$ and $\bar{\rho}_3 = (3 + \sqrt{6})/6$. On road 3 we will have a virtual shock, viz.,

$$\rho_3(x,t) = \begin{cases} \frac{3-\sqrt{6}}{6} & \text{for } x < b_3, \\ \frac{3+\sqrt{6}}{6} & \text{for } x = b_3, \end{cases} \tag{3.22}$$

while

$$\rho_j(x,t) = \begin{cases} \frac{3-\sqrt{6}}{6} & \text{for } x > \tilde{s}(t - t_0) + a_j, \\ \frac{1}{2}(1 - \sqrt{\frac{5}{6}}) & \text{for } x < \tilde{s}(t - t_0) + a_j,\ j = 2, 4, \end{cases} \tag{3.23}$$

where $\tilde{s} = \sqrt{6}/(3\sqrt{5} - 6)$. (The solutions (3.22) and (3.23) are only valid for $t > t_0$ until the shock on road 2 hits junction A (roads 2 and 3) or returns to B (road 4). After that, the solution will have to be recalculated.)

We see that the difference in flux carried by the shock will be halved after the collision at B, one shock propagating out of the system on road 4, and one shock on road 2. The shock on road 2 will, after some time, hit junction A, but since road 1 is empty, just pass this junction unchanged. Thus the system will gradually empty; each time the shock on road 2 hits junction B, it will halve its strength in γ and also move faster. However, the network will not be completely empty in finite time!

REFERENCES

[1] T. BASS, *Road to ruin*, Discover, 13 (1992), pp. 56–61.

[2] J. H. BICK AND G. F. NEWELL, *A continuum model for two-directional traffic flow*, Quart. Appl. Math., 18 (1961), pp. 191–204.

[3] D. BRAESS, *Über ein Paradoxon aus der Verkehrsplanung*, Unternehmensforschung, 12 (1968), pp. 258–268.

[4] C. M. DAFERMOS, *Polygonal approximations of solutions of the initial value problem for a conservation law*, J. Math. Anal. Appl., 38 (1972), pp. 33–41.

[5] S. DAFERMOS AND A. NAGURNEY, *On some traffic equilibrium theory paradoxes*, Transportation Res. Part B, 18 (1984), pp. 101–110.

[6] S. C. DE, *Kinematic wave theory of bottlenecks of varying capacity*, Proc. Cambridge Phil. Soc., 52 (1956), pp. 564–572.

[7] D. C. GAZIS, ED., *Traffic Science*, John Wiley, New York, 1974.

[8] H. GREENBERG, *An analysis of traffic flow*, Oper. Res., 7 (1959), pp. 79–85.

[9] R. HABERMAN, *Mathematical Models*, Prentice–Hall, Englewood Cliffs, NJ, 1977.

[10] H. HOLDEN, *On the Riemann problem of a prototype of a mixed type conservation law*, Comm. Pure Appl. Math., 40 (1987), pp. 229–264.
[11] H. HOLDEN AND L. HOLDEN, *On scalar conservation laws in one-dimension*, in Ideas and Methods in Mathematical Analysis, Stochastics, and Applications, S. Albeverio, et al., eds., Cambridge University Press, Cambridge, 1992, pp. 480–509.
[12] H. HOLDEN, L. HOLDEN, AND R. HØEGH-KROHN, *A numerical method for first order nonlinear scalar conservation laws in one–dimension*, Comput. Math. Appl., 15 (1988), pp. 595–602.
[13] H. HOLDEN, L. HOLDEN, AND N. H. RISEBRO, *Some qualitative properties of* 2×2 *systems of conservation laws of mixed type*, in Nonlinear Evolution Equations That Change Type, B. L. Keyfitz and M. Shearer, eds., Springer-Verlag, New York, 1990, pp. 67–78.
[14] H. HOLDEN AND N.H. RISEBRO, *Stochastic properties of the scalar Buckley–Leverett equation*, SIAM J. Appl. Math., 51 (1991), pp. 1472–1488.
[15] ———, *A stochastic approach to conservation laws*, in Third International Conference on Hyperbolic Problems. Theory, Numerical Methods and Applications, B. Engquist and B. Gustafsson, eds., Studentlitteratur/Chartwell-Bratt, Lund-Bromley, 1991, pp. 575–587.
[16] W. LEUTZBACH, *Introduction to the Theory of Traffic Flow*, Springer-Verlag, Berlin, 1988.
[17] R. J. LEVEQUE, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, 1990.
[18] M. J. LIGHTHILL AND G. B. WHITHAM, *On kinematic waves. I. Flood movement in long rivers*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 281–316.
[19] ———, *On kinematic waves. II. Theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.
[20] A. D. MAY, *Traffic Flow Fundamentals*, Prentice–Hall, Englewood Cliffs, NJ, 1990.
[21] S. MOCHON, *An analysis of the traffic on highways with changing surface conditions*, Math. Model., 9 (1987), pp. 1–11.
[22] P. I. RICHARDS, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.
[23] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.
[24] G. B. WHITHAM, *Linear and Nonlinear Waves*, John Wiley, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 1018–1046, July 1995

011

CHARACTERIZATION OF L^p-SOLUTIONS FOR THE TWO-SCALE DILATION EQUATIONS*

KA-SING LAU† AND JIANRONG WANG†

Abstract. We give a characterization of the existence of compactly supported L^p-solutions, $1 \leq p < \infty$, for the two-scale dilation equations. For the L^2-case, the condition reduces to the determination of the spectral radius of a certain matrix in terms of the coefficients, which can be calculated through a finite step algorithm. For the other cases, we implement the characterization by the four-coefficient dilation equation and obtain some simple sufficient conditions for the existence of the solutions. The results are compared with known ones.

Key words. cascade algorithm, compactly supported L^p-solutions, dilation equation, Fourier transformation, iteration, spectral radius, wavelet

AMS subject classifications. 26A15, 26A18, 39A10, 42A05

1. Introduction.

A *two-scale dilation equation* is a functional equation of the form

$$f(x) = \sum_{n=0}^{N} c_n f(\alpha x - \beta_n), \tag{1.1}$$

where f: $\mathbf{R} \longrightarrow \mathbf{R}$ (or $\mathbf{C}$), $\alpha > 1$, $\beta_0 < \beta_1 < \cdots < \beta_N$ are real constants, and c_n are real (or complex) constants. The equation is called a *lattice* two-scale dilation equation if

$$f(x) = \sum_{n=0}^{N} c_n f(kx - n) \tag{1.2}$$

for an integer $k \geq 2$. A special case of the functional equation ($k = 3$, $N = 4$, and $c_n = 1, 2/3, 1/3, 1/3, 1$) was first studied by de Rham [dR] as an example of a continuous nowhere differentiable function. Recently this equation has attracted a lot of attention, especially for the lattice case with $k = 2$. In wavelet theory, the study of multiresolution and the search of various orthogonal, compactly supported wavelets has lead to the investigation of the existence, uniqueness, and smoothness of such continuous integrable solutions (see the work of Cohen, Colella, Daubechies, Heil, Lagarias, Lawton, Mallat and Meyer; see the survey paper [H]). The equation also plays an important role in the "subdivision schemes" and "interpolation schemes" of constructing continuous spline curves, surfaces, and fractal objects (see the work of Cavaretta, Dahmen, Deslauriers, Dubuc, Dye, Gregory, Levin, Michelli, Prautzsch; see [DL1] and [DL2] for an historical development and references).

The general two-scale (in fact multiscale) dilation equation (1.1) arises in the consideration of self-similar measures (Hutchinson [Hu]), and the singularity of the measures induced by the infinite Bernoulli convolutions. The latter has been studied

*Received by the editors October 13, 1992; accepted for publication (in revised form) December 10, 1993.

†Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

for a long time and the question is still unsettled (see the work of Erdös, Garcia, Jessen, Salem, Wintner; see [L1] and [L2] for some recent developments and remarks). In another direction, Strichartz [JRS], [Str] studied the asymptotic behavior of the Fourier transformation of such distributions, and made many interesting observations on the averages with respect to some fractional powers.

There are two major approaches to the equation: the Fourier method (the frequency domain approach) and the iteration method (the time-domain approach). The Fourier transformation converts the functional equation to the form

$$\hat{f}(\xi) = A \prod_{j=0}^{\infty} p(\alpha^{-j}\xi),$$

where $p(\xi) = \frac{1}{\alpha} \sum_{n=0}^{N} c_n e^{i\beta_n \xi}$. Using this, Daubechies and Lagarias [DL1] proved that for $\Delta = \alpha^{-1} \sum c_n$,

(i) if $|\Delta| < 1$ or $\Delta = -1$, then (1.1) has no integrable solution;

(ii) if $\Delta = 1$ then it has *at most* one nonzero integrable solution;

(iii) if $|\Delta| > 1$ and if an integrable solution f exists, then $\Delta = \alpha^m$ for some nonnegative integer m. The dilation equation obtained by replacing the coefficients $\{c_n\}$ with $\{\alpha^{-m} c_n\}$ has a nonzero integrable solution g, and for suitable choice of normalization,

$$\frac{d^m}{dx^m} g(x) = f(x) \quad \text{a.e.}$$

For an integrable solution, the above result essentially reduces the coefficients of the equation to the special case

$$\sum c_n = \alpha.$$

By using the Fourier transform of f and the Paley–Wiener theorem, it was also proved in [DL1] that f has compact support in $[0, \beta_N N/(\alpha - 1)]$. The Fourier method, however, does not give sharp criteria for the existence of L^1-solutions in terms of the coefficients $\{c_n\}$. Some partial results are given in [La] and [M].

The iteration method is restricted to the lattice case. It applies particularly well in the case of compactly supported solutions. The basic idea is to identify a given function f supported by $[0, N]$ with the vector-valued function

$$\mathbf{f}(x) = [f(x), f(x+1), \ldots, f(x+(N-1))]^t, \quad x \in [0,1],$$

and to use the right side of the dilation equation to construct two $N \times N$ matrices T_0 and T_1 (see §2 for details). A constant vector v is used as the initial condition, followed by iteration with the matrices T_0 and T_1 (the *cascade algorithm*). The limit, if the sequence converges, will be the solution of the dilation equation. Such an approach was used by Daubechies and Lagarias [DL2], and independently by Michelli and Prautzsch [MP]. It was also used by Berger and Wang [BW1], and Collela and Heil [CH1] and [CH2].

For two given matrices A_0 and A_1, Rota and Strang [RS] and Strang [S] defined the *joint spectral radius* of A_0, A_1 by

$$\hat{\rho}(A_0, A_1) = \limsup_{m \to \infty} \lambda_m(A_0, A_1),$$

where

$$\lambda_m(A_0, A_1) = \max_{|J|=m} ||A_J||^{\frac{1}{m}}$$

with $J = (j_1, \dots, j_l)$, $A_J = A_{j_1} \cdots A_{j_l}$, $j_i = 0$ or 1. A useful sufficient condition for the existence of solutions is given in [DL2] and [BW1].

THEOREM 1.1. *For $k = 2$, $\sum c_{2n} = \sum c_{2n+1} = 1$, let*

$$H = \left\{ u = [u_1, \dots, u_N]^t : \quad \sum u_i = 0 \right\}.$$

If $\hat{\rho}(T_0|_H, T_1|_H) < 1$, then the equation has a nonzero continuous integrable solution.

Colella and Heil [CH1] and [CH2] also showed that the condition is "essentially" necessary. More recently Wang [W] introduced the notion of *mean spectral radius*:

$$\bar{\rho}(T_0, T_1) = \limsup_{m \to \infty} \frac{1}{2} \left(\sum_{|J|=m} ||T_J|| \right)^{\frac{1}{m}}.$$

He proved, among other interesting results, that if $\sum c_{2n} = \sum c_{2n+1} = 1$ and $\bar{\rho}(A_0, A_1) < 1$, where

$$T_i \approx \begin{bmatrix} 1 & 0 \\ b_i & A_i \end{bmatrix}, \qquad i = 0, 1,$$

then a nonzero integrable solution exists.

This characterization in terms of the joint spectral radius, although elegant, is difficult to evaluate in practice. By using a geometric convergence consideration and a different iteration argument, Pan [P] gives a simple sufficient condition for the existence of compactly supported L^p-solutions of the functional equation (1.2) with four coefficients.

In this paper we will continue to study the existence of the compactly supported L^p-solutions of

$$f(x) = \sum_{n=0}^{N} c_n f(2x - n), \tag{1.3}$$

using the cascade iteration algorithm with the matrices T_0 and T_1. The regularity of such solutions will be dealt with in a forthcoming paper. Note that in the previous literature, one always starts with an initial condition that is, in a certain sense, quite arbitrary (for example, a spline function or $\chi_{[0,1]}$). Our fundamental observation is the following proposition.

PROPOSITION 1.2. *Suppose $1 \le p < \infty$ and $\sum c_n = 2$. Let f be a compactly supported L^p-solution of* (1.3) *and let*

$$v = \left[\int_0^1 f, \dots, \int_{N-1}^N f \right]^t.$$

Then v is an eigenvector of $(T_0 + T_1)$ corresponding to the eigenvalue 2.

It follows that we can start with the iteration algorithm on the 2-eigenvector of $T_0 + T_1$, and the convergence condition will be imposed *only* on the subspace involved with such eigenvector. This allows us to obtain sharper results. The basic theorem is as follows.

THEOREM 1.3. *Suppose $1 \le p < \infty$. Then equation* (1.3) *has a nonzero compactly supported L^p-solution if and only if there exists a 2-eigenvector v of $(T_0 + T_1)$ such that*

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J(T_0 - I)v||^p \longrightarrow 0 \quad \text{as } l \longrightarrow \infty.$$

For computational purposes we let $H(\tilde{v})$ be the subspace in $\mathbf{R}^n$ generated by the $T_J\tilde{v}$'s for all J, where $\tilde{v} = (T_0 - I)v$, and let $\{v_1, \dots, v_k\}$ be a basis; then the above condition is equivalent to the existence of an integer l such that

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J v_i||^p < 1 \tag{1.4}$$

for all v_i, $i = 1, \dots, k$.

The above results, as well as some corollaries and remarks, are proved in §2. A slight improvement of Theorem 1.3 under the condition that the coefficients satisfy the "m-sum rules" (see (2.7)) is also considered.

In §3, we consider the equation for the three-coefficient ($N = 2$) and the four-coefficient ($N = 3$) cases. For the first case we obtain a complete characterization of the compactly supported L^p-solutions. The second case is less trivial; it contains the well-known Daubechies wavelet D_4 [D], and has been studied in detail in [H] and [P]. By using the basic theorem, we are able to derive some simple criteria for such solutions to exist.

In §4 we give an improvement of Theorem 1.3 for the L^2-case. In this case, the left-hand side of (1.4) can be calculated and leads to an explicit expression of an $N \times N$ matrix W (Lemma 4.1, Proposition 4.3). Under a stronger assumption on the coefficients

$$\sum c_{2n} = \sum c_{2n+1} = 1, \tag{1.5}$$

we show that the matrix W has an eigenvalue 2; (1.4), and hence the existence of the compactly supported L^2-solution, is essentially equivalent to the fact that all other eigenvalues of W are less than 2 (Theorem 4.4 and Proposition 4.6). For the four-coefficient case we obtain a complete characterization of the existence of the compactly supported L^2-solutions (Theorem 4.8).

There are different criteria for the existence of L^2-solutions; e.g., see [E], [Her1], [Her2], and [V]. Their approach is via a Fourier method which is quite different from ours (see Remark 9 in §4).

In [CH1], Collela and Heil used (c_0, c_3) as free parameters for the four-coefficient case satisfying $c_0 + c_2 = c_1 + c_3 = 1$, and plotted different domains in $\mathbf{R}^2$ that admit or do not admit solutions. We conclude our study with an appendix for displaying our result and some other well-known results with the same kind of plots.

2. The basic theorems. Throughout this paper we will consider the compactly supported L^p-solutions, $1 \le p < \infty$, of the functional equation

$$f(x) = \sum_{n=0}^{N} c_n f(2x - n). \tag{2.1}$$

The general lattice case can be handled similarly (see (2.5)). For convenience we let $c_n = 0$ if $n \notin \{0, \dots, N\}$. Our basic assumption on the coefficients is $\sum c_n = 2$. For some cases we will also assume that $\sum c_{2n} = \sum c_{2n+1} = 1$. We will further restrict the c_n's and the function f to be real valued, though there is no difficulty in extending our method to the complex case.

It is known that if an L^1-solution exists, then it is necessarily unique, and is supported by $[0, N]$ [DL1]. This is not true if $1 < p < \infty$, since the Hilbert transformation of such solution is again an L^p-solution [H]. We will use *L^p_c-solution* to denote

the compactly supported L^p-solution. An L^p_c-solution must be integrable, and hence supported by $[0, N]$.

Formally, the solution f is obtained by taking the limit of $\mathbf{S}^n(g)$ for a suitable function g on $\mathbf{R}$, where

$$\mathbf{S}(g)(x) = \sum_{n=0}^{N} c_n g(2x - n).$$

In some previous papers [BW], [CH1], [CH2], [DL2], [MP], [W], it is found that the analysis is a lot more convenient if we convert the involved functions into vector forms and the operator $\mathbf{S}$ into a matrix operator. For this we let

$$T_0 = [c_{2i-j-1}]_{1\le i,j\le N} = \begin{bmatrix} c_0 & 0 & 0 & \dots & 0 \\ c_2 & c_1 & c_0 & \dots & 0 \\ c_4 & c_3 & c_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \dots & c_{N-1} \end{bmatrix},$$

$$T_1 = [c_{2i-j}]_{1\le i,j\le N} = \begin{bmatrix} c_1 & c_0 & 0 & \dots & 0 \\ c_3 & c_2 & c_1 & \dots & 0 \\ c_5 & c_4 & c_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \dots & c_N \end{bmatrix}.$$

For any g defined on $\mathbf{R}$ vanishing outside $[0, N]$, we decompose g into N pieces and form a vector function as follows: let $g_i(x) = g(x + i)\chi_{[0,1)}, i = 0, 1, \dots N - 1$, and define a vector function $\mathbf{\Phi}(g) = \mathbf{g} : \mathbf{R} \longrightarrow \mathbf{R}^N$ by

$$\mathbf{g}(x) = \begin{cases} [g_0(x), g_1(x), \dots, g_{N-1}(x)]^t & \text{if } x \in [0, 1), \\ 0 & \text{if } x \notin [0, 1). \end{cases}$$

Here we use v^t to denote the transpose of a vector v. Let $||\cdot||$ be any fixed norm on $\mathbf{R}^N$ and define, as usual, $||\mathbf{g}|| = ||\mathbf{g}||_{L^p} = (\int_0^1 ||\mathbf{g}(x)||^p dx)^{1/p}$, so $g \in L^p[0, N]$ if and only if $\mathbf{g} \in L^p([0, 1], \mathbf{R}^N)$. Note that if we take the l^p-norm on $\mathbf{R}^N$, then $||\mathbf{g}||_{L^p} = ||g||_{L^p}$.

Let $\mathbf{T}$ be an operator defined on the vector-valued functions $\mathbf{g}$ by

$$(\mathbf{Tg})(x) = T_0 \cdot \mathbf{g}(\phi_0^{-1}(x)) + T_1 \cdot \mathbf{g}(\phi_1^{-1}(x)),$$

where $\phi_0(x) = \frac{1}{2}x$, $\phi_1(x) = \frac{1}{2}x + \frac{1}{2}$. Equivalently, $\mathbf{T}$ is given by

$$(\mathbf{Tg})(x) = \begin{cases} T_0 \cdot \mathbf{g}(2x) & \text{if } x \in [0, \frac{1}{2}), \\ T_1 \cdot \mathbf{g}(2x - 1) & \text{if } x \in [\frac{1}{2}, 1), \end{cases}$$

and $(\mathbf{Tg})(x) = 0$ if $x \notin [0, 1)$. If we iterate the operator $\mathbf{T}$ on $\mathbf{g}$ repeatedly and obtain a formal limit $\mathbf{f}$, then $\mathbf{f}$ will satisfy

$$\mathbf{f}(x) = T_0 \cdot \mathbf{f}(\phi_0^{-1}(x)) + T_1 \cdot \mathbf{f}(\phi_1^{-1}(x)). \tag{2.2}$$

PROPOSITION 2.1. *Let f be supported by $[0, N]$, and let $\mathbf{\Phi}(f) = \mathbf{f}$ be defined as above; then*

$$\mathbf{\Phi S}(f) = \mathbf{T\Phi}(f).$$

Moreover, f is an L^p_c-solution of (2.1) *if and only if* $\mathbf{f} \in L^p([0,1], \mathbf{R}^N)$ *and* $\mathbf{f} = \mathbf{T}\mathbf{f}$, *i.e.,* $\mathbf{f}$ *satisfies equation* (2.2).

Proof. The proof of the commutativity of the operators only involves a direct computation of x in $[i, i+\frac{1}{2}]$ and $[i+\frac{1}{2}, i+1]$. The second part is a consequence of the first part, making use of the fact that if the solution has compact support, then it must be contained in $[0, N]$. □

We begin with some simple considerations of the eigenproperties of T_0 and T_1. Unless otherwise specified, eigenvector will mean *right* eigenvector. By a λ-eigenvector of a matrix M, we will mean an eigenvector of M corresponding to the eigenvalue λ. The following proposition is known.

PROPOSITION 2.2. *If* $\sum c_n = 2$, *then* 2 *is an eigenvalue of* $(T_0 + T_1)$ *with left eigenvector* $[1, \ldots, 1]$.

Furthermore, if $\sum c_{2n} = \sum c_{2n+1} = 1$, then 1 is an eigenvalue of T_0 and T_1 with $[1, \ldots, 1]$ as a left eigenvector.

Proof. We need only observe that in the matrix $(T_0 + T_1)$, each column has sum equal to 2. The proof of the second statement is similar. □

It follows that the *right* 2-eigenvector of $(T_0 + T_1)$ exists also; it will play a central role in the existence of the solution of (2.1). Let f_Δ be the average of f over an interval Δ, i.e.,

$$f_\Delta = \frac{1}{|\Delta|} \int_\Delta |f|,$$

where $|\Delta|$ is the length of Δ.

PROPOSITION 2.3. *Let f be an L^p_c-solution of* (2.1); *let* $v = [f_{[0,1]}, \ldots, f_{[N-1,N]}]^t$ *be the vector defined by the average of f on the N subintervals as indicated. Then v is a 2-eigenvector of* $(T_0 + T_1)$.

Proof. Since $\mathbf{f} = \mathbf{T}\mathbf{f}$, i.e.,

$$\mathbf{f}(x) = \begin{cases} T_0 \cdot \mathbf{f}(2x) & \text{if } x \in [0, \frac{1}{2}), \\ T_1 \cdot \mathbf{f}(2x-1) & \text{if } x \in [\frac{1}{2}, 1), \end{cases}$$

when we integrate the expression over $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$ separately, we have

$$\begin{bmatrix} f_{[0,\frac{1}{2}]} \\ \vdots \\ f_{[N-1,N-\frac{1}{2}]} \end{bmatrix} = T_0 v, \qquad \begin{bmatrix} f_{[\frac{1}{2},1]} \\ \vdots \\ f_{[N-\frac{1}{2},N]} \end{bmatrix} = T_1 v.$$

On the other hand, note that on each interval $[i, i+1]$, the average satisfies

$$f_{[i,i+\frac{1}{2}]} + f_{[i+\frac{1}{2},i+1]} = 2 f_{[i,i+1]};$$

hence we conclude that $(T_0 + T_1)v = 2v$. □

We will show, under suitable conditions, that the 2-eigenvector of $(T_0 + T_1)$ actually defines a step function that generates the solution of (2.1). This is done by iterating with the operator $\mathbf{T}$, and is itself the average vector of the solution. For this purpose, we need to introduce some notation for the indices: For any $k \geq 1$, let

$$J = (j_1, \ldots, j_k), \quad \text{where } j_i = 0 \text{ or } 1, \quad i = 1, 2, \ldots, k,$$

and set $J = \emptyset$ if $k = 0$ for convenience; we will use $|J|$ to denote the length of J, and let

$$\Lambda = \{J : \quad |J| = k, \quad k = 0, 1, 2, \ldots\}$$

denote the class of indices. For $J, J' \in \Lambda$, we let $(J, J') = (j_1, \ldots, j_k, j'_1, \ldots, j'_{k'})$. Let I be the interval $[0,1)$; I_J will denote the dyadic interval $\phi_{j_1} \circ \phi_{j_2} \cdots \circ \phi_{j_k}([0,1))$. For example, $I_0 = [0, \frac{1}{2})$, $I_1 = [\frac{1}{2}, 1)$, and $I_J = I_{(j_1,\ldots,j_k)} = [a, b)$, where

$$a = \frac{j_1}{2} + \frac{j_2}{2^2} + \cdots + \frac{j_k}{2^k}, \qquad b = a + \frac{1}{2^k}.$$

It follows that $I_{(J,0)} \cup I_{(J,1)} = I_J$ and $I_{(J,J')} \subseteq I_J$ for any $J, J' \in \Lambda$. The matrix T_J represents the product $T_{j_1} \ldots T_{j_k}$ and $T_\emptyset$ is the identity matrix.

LEMMA 2.4. *Let* $\mathbf{f}_0(x) = v$ *for* $x \in [0,1)$, *and* $\mathbf{f}_{k+1} = \mathbf{T}\mathbf{f}_k$, $k = 0, 1, \ldots$; *then* $\mathbf{f}_k(x) = T_J v$ *for each* $x \in I_J$.

Moreover, if f *is an* L^p_c*-solution of* (2.1) *and* v *is the average vector of* f *defined in Proposition* 2.3, *then*

$$\mathbf{f}_k(x) = T_J v = [f_{I_J}, f_{(I_J+1)}, \ldots, f_{(I_J+N-1)}]^t,$$

where $(I_J + j)$ *is the interval* $\{x + j : x \in I_J\}$. *Also,* $\mathbf{f}_k \longrightarrow \mathbf{f} = \mathbf{\Phi}(f)$ *in* $L^p([0,1], \mathbf{R}^N)$.

Proof. We will use induction to show that $\mathbf{f}_k(x) = T_J v$ for $x \in I_J$ with $|J| = k$. Suppose that $\mathbf{f}_k(x) = T_J v$ for $x \in I_J$. Let $x \in I_{(0,J)} = \phi_0(I_J)$; then $\phi_0^{-1}(x) = 2x \in I_J$ and

$$\mathbf{f}_{k+1}(x) = \mathbf{T}(\mathbf{f}_k(x)) = T_0 \cdot \mathbf{f}_k(2x) = T_0 T_J v = T_{(0,J)} v.$$

Similarly, if $x \in I_{(1,J)}$, then $\mathbf{f}_{k+1}(x) = T_{(1,J)} v$.

Let $\mathbf{f} = \mathbf{\Phi}(f)$; then $\mathbf{f} = \mathbf{T}\mathbf{f}$ and $\mathbf{f}(x) = T_J \mathbf{f}(\phi_J^{-1}(x))$ for $x \in I_J$. Integrating this over the interval I_J, we obtain

$$[f_{I_J}, \ldots, f_{I_J+N-1}]^t = T_J v.$$

The fact that $\mathbf{f}_k \longrightarrow \mathbf{f}$ in $L^p([0,1], \mathbf{R}^N)$ follows by a proposition in [R, p. 129]. □

LEMMA 2.5. *Let* v *be a* 2*-eigenvector of* $(T_0 + T_1)$, *and let* $\mathbf{f}_k$ *be defined as above*; *then for each* k,

$$\int_{[0,1]} \mathbf{f}_k(x)dx = v. \tag{2.3}$$

Proof. Equation (2.3) follows from the following induction argument:

$$\begin{aligned}
\int_{[0,1]} \mathbf{f}_{k+1}(x)dx &= \int_{[0,\frac{1}{2}]} T_0 \cdot \mathbf{f}_k(2x)dx + \int_{[\frac{1}{2},1]} T_1 \cdot \mathbf{f}_k(2x-1)dx \\
&= \frac{1}{2}\left(T_0 \int_{[0,1]} \mathbf{f}_k(x)dx + T_1 \int_{[0,1]} \mathbf{f}_k(x)dx \right) \\
&= \frac{1}{2}(T_0 + T_1) \int_{[0,1]} \mathbf{f}_k(x)dx \\
&= \frac{1}{2}(T_0 + T_1)v = v. \qquad \square
\end{aligned}$$

For any 2-eigenvector v of $(T_0 + T_1)$, we have

$$(T_0 - I)v = -(T_1 - I)v.$$

Let $\tilde{v} = (T_0 - I)v$ and $H(\tilde{v})$ be the subspace in $\mathbf{R}^N$ spanned by

$$\{T_J\tilde{v} : J \in \Lambda\}.$$

THEOREM 2.6. *For $1 \le p < \infty$, the following are equivalent*: (i) *equation* (2.1) *has a nonzero L_c^p-solution*; (ii) *there exists a 2-eigenvector v of $(T_0 + T_1)$ satisfying*

$$\lim_{l \longrightarrow \infty} \frac{1}{2^l} \sum_{|J|=l} ||T_J\tilde{v}||^p = 0;$$

(iii) *there exists a 2-eigenvector v of $(T_0 + T_1)$ such that there exists an integer $l \ge 1$ such that*

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J u||^p < 1 \quad \textit{for all } u \in H(\tilde{v}), \quad ||u|| \le 1. \tag{2.4}$$

Proof. Let $\mathbf{f}_0 = v$ and $\mathbf{f}_{n+1} = \mathbf{T}\mathbf{f}_n$. By Lemma 2.4, for $x \in I_J$ and $|J| = n$, $\mathbf{f}_n(x) = T_J v$. Let $\mathbf{g}_n = \mathbf{f}_{n+1} - \mathbf{f}_n$; then $\mathbf{f}_{n+1} = \mathbf{f}_0 + \mathbf{g}_0 + \cdots + \mathbf{g}_n$, where

$$\mathbf{g}_n(x) = \begin{cases} T_{(J,0)}v - T_J v = T_J\tilde{v} & \text{if } x \in I_{(J,0)}, \\ T_{(J,1)}v - T_J v = -T_J\tilde{v} & \text{if } x \in I_{(J,1)}, \end{cases}$$

and

$$||\mathbf{g}_n||^p = \frac{1}{2^n} \sum_{|J|=n} ||T_J\tilde{v}||^p.$$

Since (i) implies that $||\mathbf{g}_n||$ converges to zero, (ii) follows immediately.

To prove that (ii) implies (iii), we note that $H(\tilde{v})$ is finite dimensional and has a finite basis of $T_{J'}\tilde{v}$'s. Let $u = T_{J'}\tilde{v}$ with $|J'| = k$; then

$$\frac{1}{2^n} \sum_{|J|=n} ||T_J u||^p = \frac{1}{2^n} \sum_{|J|=n} ||T_J T_{J'}\tilde{v}||^p \le 2^k \frac{1}{2^{n+k}} \sum_{|J|=n+k} ||T_J\tilde{v}||^p \longrightarrow 0$$

as $n \longrightarrow \infty$, and the convergence is uniform for all $||u|| \le 1$. Hence (2.4) follows by taking $l = n$ for n sufficiently large.

Now assume (iii) holds. Since $H(\tilde{v})$ is finite dimensional, there is a constant $0 < c < 1$ such that for any $u \in H(\tilde{v})$,

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J u||^p < c||u||^p.$$

For any $|J'| = n$, let $u = T_{J'}\tilde{v} \in H(\tilde{v})$; then

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J T_{J'}\tilde{v}||^p < c||T_{J'}\tilde{v}||^p.$$

Summing over all $|J'| = n$, we have

$$\frac{1}{2^{l+n}} \sum_{|J|=l+n} ||T_J\tilde{v}||^p = \frac{1}{2^{l+n}} \sum_{|J|=l} \sum_{|J'|=n} ||T_J T_{J'}\tilde{v}||^p < \frac{c}{2^n} \sum_{|J'|=n} ||T_{J'}\tilde{v}||^p.$$

It follows from the expression of $||\mathbf{g}_n||$ given above that

$$||\mathbf{g}_{n+l}||^p < c||\mathbf{g}_n||^p.$$

For each fixed n, $\{||\mathbf{g}_{n+kl}||\}_{k=1}^{\infty}$ is dominated by a geometric series, hence $\mathbf{f}_{n+1} = \mathbf{f}_0 + \mathbf{g}_0 + \cdots + \mathbf{g}_n$ converges in L^p. The limit $\mathbf{f}$ is nonzero by Lemma 2.5, and so by Proposition 2.1, (i) follows. □

Remark 1. If

$$\frac{1}{2^l}\sum_{|J|=l}||T_J|_{H(\tilde{v})}||^p < 1,$$

then (2.4) is satisfied. Hence, if the joint spectral radius [BW1], [DL1] or the mean spectral radius [W] of $\{T_0|_{H(\tilde{v})}, T_1|_{H(\tilde{v})}\}$ is less than 1, then a nonzero L^1-solution exists.

Remark 2. If condition (2.4) is satisfied for one particular norm on $\mathbf{R}^N$, then it will be satisfied for all the (equivalent) norms (the integer l will depend on the choice of norms). This follows directly from Theorem 2.6 (ii).

Also, condition (2.4) can be replaced by the following slightly simpler condition:

$$\frac{1}{2^l}\sum_{|J|=l}||T_J u_i||^p < 1,$$

where $\{u_1, \ldots, u_k\}$ is a basis of $H(\tilde{v})$. To see this, we define a norm on $\mathbf{R}^N$ such that its restriction on $H(\tilde{v})$ is the l^p-norm given by

$$|||u|||^p = \sum_{i=1}^{k}|\alpha_i|^p, \qquad \text{where } u = \sum_{i=1}^{k}\alpha_i u_i.$$

Let $u = \sum_{i=1}^{k}\alpha_i u_i \in H(\tilde{v})$; then

$$\frac{1}{2^l}\sum_{|J|=l}|||T_J u|||^p \le \frac{1}{2^l}\sum_{|J|=l}\sum_{i=1}^{k}|\alpha_i|^p|||T_J u_i|||^p < \sum_{i=1}^{k}|\alpha_i|^p = |||u|||^p,$$

which implies (2.4).

For computational purposes it would be interesting to know the optimal choice of a bound of the integer l in condition (2.4), in particular, when the norm on $\mathbf{R}^N$ is the l^p-norm.

Remark 3. In [DL1, Thm. 3.1 and Rem. 1], it is proved that if $\sum c_n = 2$ and a nonzero compactly supported tempered distributional solution f exists, then the Fourier transform of f must have the form

$$\hat{f}(\xi) = A\prod_{k=1}^{\infty} m_0(2^{-k}\xi),$$

where $m_0(\xi) = \frac{1}{2}\sum_{n=0}^{N-1} c_n e^{in\xi}$. Moreover, if f is integrable, then $A = \int f(x)dx$. It follows that f is unique up to a multiplicative constant. By Proposition 2.3, the above v equals $[f_{[0,1]}, \ldots, f_{[N-1,N]}]^t$, so that the 2-eigenvector satisfying (2.4) is unique. Also, it follows from the expression of A that

$$\sum_{n=0}^{N} v_i = \int f(x)dx \neq 0,$$

hence $v \notin H = \{u : \sum_{i=0}^{N} u_i = 0\}$.

We are not able to prove these two facts without using the Fourier transform. Nevertheless, we have the following result, whose negation is useful in proving the nonexistence of solutions.

COROLLARY 2.7. *Under the same hypotheses of Theorem* 2.6, *assume that the solution f exists; then $v \notin H(\tilde{v})$, and the dimension of $H(\tilde{v})$ is $\leq N-1$.*

Proof. By Theorem 2.6 (ii),

$$\frac{1}{2^n} \sum_{|J|=n} ||T_J u||^p \longrightarrow 0 \quad \text{for any } u \in H(\tilde{v}).$$

It follows that if $v \in H(\tilde{v})$, then

$$||v||^p = \frac{1}{2^n} ||(T_0 + T_1)^n v||^p \leq \frac{1}{2^n} \sum_{|J|=n} ||T_J v||^p \longrightarrow 0$$

as $n \longrightarrow \infty$. This contradicts $v \neq 0$. □

Remark 4. In the construction of the solution f, if we start the iteration from a vector other than the 2-eigenvector v, then the process may not converge, or may converge to the zero function. For example, consider

$$f(x) = f(2x) + f(2x-2).$$

In this case $c_0 = 1, c_1 = 0, c_2 = 1$, and

$$T_0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad T_1 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

The 2-eigenvector of $(T_0 + T_1)$ is $v = [1,1]^t$ and $H(\tilde{v}) = 0$, hence condition (2.4) is satisfied. The (normalized) solution f is the characteristic function of the interval $[0,2]$. However, if we start with the vector $[1,0]^t$, then the iteration with $\mathbf{T}$ will not converge.

Nevertheless, we are still able to choose a large class of vectors that can serve as initial values.

COROLLARY 2.8. *Suppose $\sum c_n = 2$. Let $w = [w_1, w_2, \ldots, w_N]^t$ be a vector in $\mathbf{R}^N$ and $H'(w)$ be the subspace spanned by*

$$\{T_J(T_i - I)w : \quad J \in \Lambda, \quad i = 0,1\}.$$

Suppose $w \notin H'(w)$ or $\sum w_i \neq 0$, and suppose

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J u||^p < 1 \quad \textit{for all } u \in H'(w), \quad ||u|| \leq 1;$$

then (2.1) *has a nonzero L_c^p-solution.*

Proof. As in the first part of the proof of Theorem 2.6, we define $\mathbf{f}_0(x) = w$ for all $x \in [0,1]$ and $\mathbf{f}_{k+1} = \mathbf{T}\mathbf{f}_k$; then $\mathbf{f}_k$ will converge in $L^p([0,1], \mathbf{R}^N)$ and the limiting function $\mathbf{f}$ will satisfy equation (2.2).

We still need to show that $\mathbf{f} \neq 0$. If $w \notin H'(w)$, then from the proof of Theorem 2.6 we have

$$\mathbf{f} = \lim_{n\to\infty} \mathbf{f}_k = \mathbf{f}_0 + \lim_{n\to\infty} \sum_{j=0}^{n} \mathbf{g}_j$$

with $\mathbf{f}_0(x) \notin H'(w)$ and $\mathbf{g}_n(x) \in H'(w)$, so $\mathbf{f}(x) \notin H'(w)$ and $\mathbf{f}$ is nonzero.

To prove the second case, we assume that $\sum w_i \neq 0$; then by Proposition 2.2, $e = [1, \ldots, 1]$ is a left 2-eigenvector of $(T_0 + T_1)$, so

$$\begin{aligned} & e \cdot \int_{[0,1]} \mathbf{f}_{k+1}(x)dx \\ &= e \cdot \int_{[0,\frac{1}{2}]} T_0 \cdot \mathbf{f}_k(2x)dx + e \cdot \int_{[\frac{1}{2},1]} T_1 \cdot \mathbf{f}_k(2x-1)dx \\ &= \frac{1}{2} e \cdot \left(T_0 \int_{[0,1]} \mathbf{f}_k(x)dx + T_1 \int_{[0,1]} \mathbf{f}_k(x)dx \right) \\ &= \frac{1}{2} e \cdot (T_0 + T_1) \int_{[0,1]} \mathbf{f}_k(x)dx = e \cdot \int_{[0,1]} \mathbf{f}_k(x)dx. \end{aligned}$$

Repeating this argument, we have

$$e \cdot \int_{[0,1]} \mathbf{f}_{k+1}(x)dx = \cdots = e \cdot \mathbf{f}_0(x) = e \cdot w = \sum w_i \neq 0.$$

This implies that $\int_{[0,1]} \mathbf{f} \neq 0$, and the proof is complete. □

Remark 5. Let D_l^p be the set of $(c_0, \ldots, c_N)$ for which (2.4) holds; then
(i) $D_l^p \subseteq D_{2l}^p$, and
(ii) $D_l^p \subseteq \cup_k D_{2^k}^p$ for any l.
Indeed, if (2.4) holds for some l, then

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J u||^p < ||u||^p \quad \text{for all} \quad u \in H(\tilde{v}).$$

Since $T_J u \in H(\tilde{v})$ if $u \in H(\tilde{v})$, we have

$$\frac{1}{2^{2l}} \sum_{|J|=2l} ||T_J u||^p \leq \frac{1}{2^l} \frac{1}{2^l} \sum_{|J'|=l} \sum_{|J''|=l} ||T_{J'} T_{J''} u||^p < \frac{1}{2^l} \sum_{|J''|=l} ||T_{J''} u||^p \leq ||u||^p.$$

To show (ii), let c be a number $0 < c < 1$ such that

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J u||^p < c||u||^p \quad \text{for all} \quad u \in H(\tilde{v})$$

holds. If $|J| = 2^k$, we write $J = (J_1, \ldots, J_m, J')$, where $|J_i| = l$ and $|J'| < l$; then

$$\begin{aligned} \frac{1}{2^{2^k}} \sum_{|J|=2^k} ||T_J u||^p &\leq \frac{1}{2^{|J'|}} \left(\frac{1}{2^l} \right)^m \sum_{J_1=l} \cdots \sum_{J_m=l} \sum_{J'} ||T_{J_1} \cdots T_{J_m} T_{J'} u||^p \\ &< c^m \frac{1}{2^{|J'|}} \sum_{J'} ||T_{J'} u||^p, \end{aligned}$$

which is less than 1 for $||u|| \leq 1$ if k (hence m) is large enough.

COROLLARY 2.9. *Equation* (2.1) *has a nonzero* L_c^p*-solution if and only if* $(c_0, \ldots, c_N) \in \cup_{k=1}^{\infty} D_{2^k}^p$.

Remark 6. One can also consider the functional equation

$$f(x) = \sum c_n f(kx - \beta n) \tag{2.5}$$

for some integer $k > 1$ and constant $\beta \neq 0$. Note that the L_c^p-solution will be supported by $[0, \beta N/(k-1)]$ [DL1]. Theorem 2.6 still holds with minor modifications of the proof. The matrices for the cascade algorithm will be

$$T_m = [c_{ki+m-j}], \quad 0 \leq i, j \leq N-1,$$

for $m = 0, \ldots, k-1$. If we define $\phi_m(x) = \frac{x}{k} + \frac{m\beta}{k}$, $m = 0, \ldots, k-1$, the vector form of equation (2.5) becomes

$$\mathbf{f}(x) = \sum_{m=0}^{k-1} T_m \mathbf{f}(\phi_m^{-1}(x)),$$

and the proof follows as above.

THEOREM 2.10. *Suppose* $\sum_{n=0}^{N} c_n = k$ *and* $1 \leq p < \infty$. *Then equation* (2.5) *has a nonzero* L_c^p*-solution if and only if there exists a* k*-eigenvector* v *of* $\sum_m T_m$ *satisfying the following: there exists an integer* $l \geq 1$ *such that*

$$\frac{1}{k^l} \sum_{j_i = 0, \ldots, k-1} ||T_{j_1} \cdots T_{j_l} u||^p < 1$$

for all vectors u, *with* $||u|| \leq 1$, *in the smallest subspace containing* $(T_m - I)v$ *and which is invariant under* T_m ,$m = 0, \ldots, k-1$.

To conclude this section we consider some special cases of Theorem 2.6. First, we assume that $\sum c_{2n} = \sum c_{2n+1} = 1$. This is a necessary condition for the solution to be the scaling function of a wavelet that defines a multiresolution (see [DL1], [CH2]). By Proposition 2.2, we know that $e = [1, \ldots, 1]$ is a common left 1-eigenvector of the two matrices T_0 and T_1. Since $\mathbf{f}_n(x) = T_J v$ if $x \in I_J = \phi_J([0,1))$,

$$e \cdot \mathbf{f}_n(x) = e \cdot T_J v = e \cdot v = \sum v_i.$$

Hence $e \cdot \mathbf{f}(x)$ equals the constant $\sum v_i$ for almost all $x \in [0,1]$; that is,

$$\sum_{n=0}^{N-1} f(x+n) = \sum v_i \quad \text{for almost all } x \in [0,1],$$

and

$$\int_0^N f(x)dx = \sum_{n=0}^{N-1} \int_0^1 f(x+n)dx = \sum v_i,$$

which is not zero as we mentioned in Remark 3.

Let H be the hyperplane of $\mathbf{R}^N$ defined by

$$H = \left\{ [u_0, \ldots, u_{N-1}]^t : \quad \sum u_j = 0 \right\};$$

then H is invariant under T_0 and T_1. For any vector $v \in \mathbf{R}^N$, we always have $(T_0 - I)v$, $(T_1 - I)v \in H$ (since the sum of the coordinates of $(T_0 - I)v$ equals $e \cdot (T_0 - I)v = 0$). Hence $H(\tilde{v}) \subseteq H$.

COROLLARY 2.11. *Suppose that* $\sum c_{2n} = \sum c_{2n+1} = 1$. *If there exists an integer* $l \geq 1$ *such that*

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J u||^p < 1 \quad \text{for all } u \in H, \quad ||u|| \leq 1, \tag{2.6}$$

then equation (2.1) *has nonzero* L^p_c*-solutions.*

Assuming $\sum c_i = 2$, the condition $\sum c_{2n} = \sum c_{2n+1} = 1$ is equivalent to $\sum(-1)^n c_n = 0$. More generally, we can consider the *m-sum rules*; that is,

$$\sum_{n=0}^{N} (-1)^n n^j c_n = 0 \quad \text{for } j = 0, 1, \ldots, m. \tag{2.7}$$

The m-sum rule is used to ensure higher order of regularity (see [DL2]). It is known that there is a matrix B such that

$$BT_0B^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ * & 1/2 & & \vdots & \vdots & & \vdots \\ \vdots & & & 0 & 0 & & 0 \\ * & \cdots & * & 1/2^m & 0 & \cdots & 0 \\ * & * & \cdots & * & * & \cdots & * \\ \vdots & & & \vdots & \vdots & & \vdots \\ * & * & \cdots & * & * & \cdots & * \end{bmatrix},$$

$$BT_1B^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ * & 1/2 & & \vdots & \vdots & & \vdots \\ \vdots & & & 0 & 0 & & 0 \\ * & \cdots & * & 1/2^m & 0 & \cdots & 0 \\ * & * & \cdots & * & * & \cdots & * \\ \vdots & & & \vdots & \vdots & & \vdots \\ * & * & \cdots & * & * & \cdots & * \end{bmatrix}.$$

The matrix B can be orthonormalized by the Gram–Schmidt process; the first row of B is the vector $[1/\sqrt{N}, \ldots, 1/\sqrt{N}]$ and the first $(m+1)$ rows of B are linear combinations of vectors

$$[1^j, 2^j, \ldots, N^j], \qquad j = 0, 1, \ldots, m.$$

Let $H' = \{[0, u_1, \ldots, u_{N-1}]^t\}$, $H'_m \subseteq H'$ be the subspace of vectors whose first $(m+1)$ components are zero. Then $H = B^{-1}H'$, and $H_m := B^{-1}H'_m$ is actually the following subspace:

$$H_m = \left\{ [u_0, \ldots, u_{N-1}]^t : \sum_{n=0}^{N-1} n^j u_n = 0, j = 0, 1, \ldots, m \right\}$$

(here $0^0 = 1$).

COROLLARY 2.12. *Suppose* $\sum c_n = 2$ *and the m-sum rules hold. If there exists an integer* $l \geq 1$ *such that*

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J u||^p < 1 \quad \text{for all } u \in H_m, \quad ||u|| \leq 1,$$

then equation (2.1) *has a nonzero* L_c^p*-solution.*

Proof. For $j = 1, \ldots, m$, let $v_j' = [0, \ldots, 0, 1, 0, \ldots, 0]^t$ be the vectors whose jth component is 1. Let $v_j = B^{-1}v_j'$; then $v_j' \in H'$, $v_j \in H$, and

$$BT_0 v_j = BT_0 B^{-1} \cdot Bv_j = BT_0 B^{-1} v_j' = \frac{1}{2^j} v_j' + w_j'$$

for some $w_j' \in H_m'$, so $T_0 v_j = (1/2^j)v_j + w_{0,j}$ for some $w_{0,j} \in H_m$. Similarly, $T_1 v_j = (1/2^j)v_j + w_{1,j}$ for some $w_{1,j} \in H_m$.

Note that $\{v_1, \ldots, v_m\}$ and H_m span H. By Corollary 2.11 and Remark 2, we need only show that for each $j = 1, \ldots, m$ there exists k such that

$$s_k := \frac{1}{2^k} \sum_{|J|=k} ||T_J v_j||^p < 1.$$

Let $c_p = 2^{p-1}$; then for the usual l^p norm we have $||u+v||^p \leq c_p(||u||^p + ||v||^p)$ for any vectors u and v. For any $j = 1, \ldots, m$ and $\eta = 1/4c_p$, by assumption there is an integer l such that

$$\frac{1}{2^l} \sum_{|J|=l} ||T_J w_{i,j}||^p < \eta \quad \text{for } i = 0, 1.$$

Then

$$\begin{aligned} s_n &= \frac{1}{2^n} \left[\sum_{|J|=n-1} \left\| T_J \left(\frac{1}{2^j} v_j + w_{0,j} \right) \right\|^p + \sum_{|J|=n-1} \left\| T_J \left(\frac{1}{2^j} v_j + w_{1,j} \right) \right\|^p \right] \\ &\leq \frac{1}{2^n} \left[\frac{2c_p}{2^{jp}} \sum_{|J|=n-1} ||T_J v_j||^p + c_p \sum_{|J|=n-1} ||T_J w_{0,j}||^p + c_p \sum_{|J|=n-1} ||T_J w_{1,j}||^p \right] \\ &< \frac{c_p}{2^{jp}} s_{n-1} + \eta c_p \leq \frac{1}{2} s_{n-1} + \frac{1}{4}. \end{aligned}$$

Hence

$$s_{n+l} < \frac{1}{2^n} s_l + \left(\frac{1}{2^{n-1}} + \frac{1}{2^{n-2}} + \cdots + 1 \right) \frac{1}{4} \leq \frac{1}{2} + \frac{1}{2} = 1$$

for sufficiently large n. □

3. Special cases: $N \leq 3$. The simplest nontrivial 2-dilation equation occurs when $N = 2$, i.e.,

$$f(x) = c_0 f(2x) + c_1 f(2x-1) + c_2 f(2x-2), \tag{3.1}$$

where $c_0 + c_1 + c_2 = 2$.

THEOREM 3.1. *For $1 \le p < \infty$, equation (3.1) has a (nonzero) L^p_c-solution if and only if either $c_1 = 1$ and*

$$\frac{1}{2}(|c_0|^p + |1 - c_0|^p) < 1,$$

or $c_0 = c_2 = 1$. In the later case $f = c\chi_{[0,2)}$.

Proof. We will use the l^p-norm on $\mathbf{R}^2$. Note that

$$T_0 = \begin{bmatrix} c_0 & 0 \\ c_2 & c_1 \end{bmatrix}, \quad T_1 = \begin{bmatrix} c_1 & c_0 \\ 0 & c_2 \end{bmatrix}, \quad \text{and } T_0 + T_1 = \begin{bmatrix} c_0 + c_1 & c_0 \\ c_2 & c_1 + c_2 \end{bmatrix}.$$

If $(c_0, c_2) = (0,0)$, then $(T_0 + T_1) = 2I$. Any nonzero vector $v = [x, y]^t$ will be a 2-eigenvector. It is a direct calculation that $v \in H(\tilde{v})$ and, by Corollary 2.7, no nonzero L^p_c-solution exists.

We assume that $(c_0, c_2) \neq (0,0)$; the 2-eigenvector of $(T_0 + T_1)$ is $v = [c_0, c_2]^t$, so that

$$\tilde{v} = (T_0 - I)v = \begin{bmatrix} c_0(c_0 - 1) \\ c_2(1 - c_2) \end{bmatrix}. \tag{3.2}$$

For an L^p_c-solution to exist, $H(\tilde{v})$ can only be $\{0\}$ or one-dimensional (Corollary 2.7).

In the first case, $\tilde{v} = 0$, condition (2.4) is automatically satisfied. The only possible cases are

$$(c_0, c_2) = (1,1),\ (0,1),\ \text{or } (1,0),$$

and the (normalized) solutions are given by $f(x) = \chi_{[0,2)}$, $\chi_{[1,2)}$, or $\chi_{[0,1)}$, respectively.

In the second case, $\tilde{v} \neq 0$. Since $H(\tilde{v})$ is invariant under T_0 and T_1, $T_0\tilde{v} = c\tilde{v}$ for some c. Expression (3.2) yields the following cases (excluding those considered above):

(a) $c_i = 0$ for $i = 0$ or 2. In this case $v \in H(\tilde{v})$ and Corollary 2.7 implies that (3.1) has no L^p_c-solution.

(b) $c_i = 1$ for $i = 0$ or 2. In this case a direct calculation shows that $T_0\tilde{v}$, $T_1\tilde{v}$ are independent. Hence $H(\tilde{v})$ is two-dimensional and by Corollary 2.7 no L^p_c-solution exists.

(c) $c_i \neq 0, 1$ for $i = 0$ and 2. By equating (3.2) and

$$T_0\tilde{v} = \begin{bmatrix} c_0^2(c_0 - 1) \\ c_2(c_0^2 + c_2^2 + c_0c_2 - 2c_0 - 3c_2 + 2) \end{bmatrix} \tag{3.3}$$

with $T_0\tilde{v} = c\tilde{v}$, we have $c = c_0$, so that by (3.2) and (3.3),

$$(c_0^2 + c_2^2 + c_0c_2 - 2c_0 - 3c_2 + 2) = c_0(1 - c_2);$$

that is

$$(c_0 + c_2 - 2)(c_0 + c_2 - 1) = 0.$$

Hence, either (i) or (ii) below holds.

(i) $c_0 + c_2 = 2$. In this case $v = [c_0, 2 - c_0]^t$ and $\tilde{v} = (c_0 - 1)v$. Once again $v \in H(\tilde{v})$ and no L^p_c-solution exists.

(ii) $c_0 + c_2 = 1$. In this case a direct calculation shows that $T_0\tilde{v} = c_0\tilde{v}$, $T_1\tilde{v} = c_2\tilde{v}$. By Theorem 2.6, equation (3.1) has an L^p_c-solution if and only if there exists an integer $l \geq 1$ such that

$$\frac{1}{2^l}(|c_0|^p + |c_2|^p)^l ||\tilde{v}||^p = \frac{1}{2^l} \sum_{|J|=l} ||T_J\tilde{v}||^p < ||\tilde{v}||^p.$$

This is equivalent to

$$\frac{1}{2}(|c_0|^p + |1 - c_0|^p) < 1.$$

The theorem follows by summarizing all the cases. □

It follows directly from the theorem that if $c_0 + c_2 = 1$ and if

(a) $c_0 \in (-\frac{1}{2}, \frac{3}{2})$, then an L^1_c-solution exists;

(b) $c_0 \in (\frac{1-\sqrt{3}}{2}, \frac{1+\sqrt{3}}{2})$, then an L^2_c-solution exists;

(c) $c_0 \in (0, 1)$, then an L^p_c-solution exists for all $1 \leq p < \infty$.

The conditions are also necessary except for $f = \chi_{[0,2)}$. We remark that in [W] it is proved that if $c_0 + c_2 = 1$, then equation (3.1) has a continuous solution if and only if $c_0 \in (0, 1)$, which is stronger than (c). Other proofs of the L^1-, L^2-cases in (a) and (b) are also known (see [P]).

We will now consider the 2-dilation equation with $N = 3$:

$$f(x) = c_0 f(2x) + c_1 f(2x - 1) + c_2 f(2x - 2) + c_3 f(2x - 3) \tag{3.4}$$

with the stronger assumption $c_0 + c_2 = c_1 + c_3 = 1$. The matrices T_0 and T_1 are given by

$$T_0 = \begin{bmatrix} c_0 & 0 & 0 \\ c_2 & c_1 & c_0 \\ 0 & c_3 & c_2 \end{bmatrix}, \quad T_1 = \begin{bmatrix} c_1 & c_0 & 0 \\ c_3 & c_2 & c_1 \\ 0 & 0 & c_3 \end{bmatrix}.$$

It is easy to show that $(T_0 + T_1)$ has 1, 2, and $(1 - c_0 - c_3)$ as eigenvalues, and the 2-eigenvector is

$$v = \begin{bmatrix} c_0(1 + c_0 - c_3) \\ (1 + c_0 - c_3)(1 - c_0 + c_3) \\ c_3(1 - c_0 + c_3) \end{bmatrix}$$

provided that $(c_0, c_3) \neq (0, -1)$ or $(-1, 0)$ (in these cases the 2-eigenvectors are given by $[1, 0, 0]^t$ and $[0, 0, 1]^t$, respectively). It follows that (excluding the two exceptional cases),

$$\tilde{v} = (T_0 - I)v = \begin{bmatrix} c_0(c_0 - 1)(1 + c_0 - c_3) \\ -c_0(c_0 - 1)(1 + c_0 - c_3) - c_3(1 - c_3)(1 - c_0 + c_3) \\ c_3(1 - c_3)(1 - c_0 + c_3) \end{bmatrix}.$$

Recall that the subspace $H(\tilde{v})$ is generated by $T_J(\tilde{v})$, $J \in \Lambda$. Under the assumption $c_0 + c_2 = c_1 + c_3 = 1$, T_J is invariant on

$$H = \{[x, y, z]^t : x + y + z = 0\},$$

and $H(\tilde{v}) \subseteq H$. For convenience we will reduce the matrices T_0 and T_1 on H by considering the first and the third coordinates of $[x, y, z]^t$ in H. This defines two matrices S_i, $i = 0, 1$, as in [CH1]. This can be seen by the following diagram:

$$\begin{array}{ccc} H & \xrightarrow{T_i} & H \\ \tau \uparrow & & \downarrow \tau^{-1}, \\ \mathbf{R}^2 & \xrightarrow{S_i} & \mathbf{R}^2 \end{array}$$

where $S_i = \tau^{-1}T_i\tau$ with $\tau : \mathbf{R}^2 \longrightarrow H$ denoting the natural isomorphism $[x, z] \longrightarrow [x, -(x+z), z]$. The explicit expression of S_i, $i = 1, 2$, is given by

$$S_0 = \begin{bmatrix} c_0 & 0 \\ -c_3 & 1-c_0-c_3 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 1-c_0-c_3 & -c_0 \\ 0 & c_3 \end{bmatrix}.$$

Slightly abusing the notation, we will still use $\tilde{v}$ and $H(\tilde{v})$ in $\mathbf{R}^2$ for the corresponding terms in H. The new $\tilde{v}$ in $\mathbf{R}^2$ is given by

$$\tilde{v} = \begin{bmatrix} c_0(c_0-1)(1+c_0-c_3) \\ c_3(1-c_3)(1-c_0+c_3) \end{bmatrix} \tag{3.5}$$

for $(c_0, c_3) \neq (0, -1)$ or $(-1, 0)$. The following theorem follows readily from Theorem 2.6.

THEOREM 3.2. *For $1 \leq p < \infty$, equation (3.4) has a nonzero L^p_c-solution if and only if there exists an integer l such that*

$$\frac{1}{2^l} \sum_{|J|=l} ||S_J u||^p < 1 \quad \textit{for all } u \in H(\tilde{v}) \quad \textit{and } ||u|| \leq 1. \tag{3.6}$$

For the degenerate case (i.e., $H(\tilde{v}) = \{0\}$ or one-dimensional), condition (3.6) can be displayed explicitly. This is shown in the following two lemmas.

LEMMA 3.3. *$H(\tilde{v}) = \{0\}$ if and only if $(c_0, c_3) \in \{(0,0), (1,0), (0,1), (1,1)\}$.*

Proof. This is a consequence of (3.5), and a direct computation of the two special cases $(c_0, c_3) = (0, -1)$ or $(-1, 0)$ (for such cases the corresponding $H(\tilde{v})$ are two dimensional). □

The solutions for these special cases can be handled easily as follows:

If $(c_0, c_3) = (0, 0)$, then the solutions are $f = c\chi_{[1,2]}$.

If $(c_0, c_3) = (1, 0)$, then the solutions are $f = c\chi_{[0,1]}$.

If $(c_0, c_3) = (0, 1)$, then the solutions are $f = c\chi_{[2,3]}$.

If $(c_0, c_3) = (1, 1)$, then the solutions are $f = c\chi_{[0,3]}$.

It is also simple to show that for the exceptional cases $(c_0, c_3) = (0, -1)$ or $(-1, 0)$, condition (3.6) is not satisfied; therefore there is no L^p_c-solution.

LEMMA 3.4. *$H(\tilde{v})$ is one-dimensional if and only if $(c_0, c_3) \notin \{(0,0), (1,0), (0,1), (1,1)\}$ and one of the following holds:*

$$c_0 = 0, \quad c_3 = 0, \quad \textit{or } 1 - c_0 - c_3 = 0.$$

Let $\tilde{v}_0 = S_0\tilde{v}$, $\tilde{v}_1 = S_1\tilde{v}$. Then for the above three cases we have

$$\begin{aligned} &\tilde{v} = c[0,1]^t, \quad \textit{and } \tilde{v}_0 = (1-c_0-c_3)\tilde{v}, \ \tilde{v}_1 = c_3\tilde{v}; \\ &\tilde{v} = c[1,0]^t, \quad \textit{and } \tilde{v}_0 = c_0\tilde{v}, \ \tilde{v}_1 = (1-c_0-c_3)\tilde{v}; \\ &\tilde{v} = c[c_0, -c_3]^t, \quad \textit{and } \tilde{v}_0 = c_0\tilde{v}, \ \tilde{v}_1 = c_3\tilde{v}, \end{aligned}$$

respectively.

Proof. The sufficiency is clear; we only prove the necessity. Assuming c_0, $c_3 \neq 0$, we want to show that $(1 - c_0 - c_3) = 0$.

Suppose $(1 - c_0 - c_3) \neq 0$. Note that S_0 has two eigenvalues c_0 and $(1 - c_0 - c_3)$ with corresponding eigenvectors $[1-2c_0-c_3, c_3]^t$ and $[0, 1]^t$, and S_1 has two eigenvalues c_3 and $(1 - c_0 - c_3)$ with corresponding eigenvectors $[c_0, 1 - c_0 - 2c_3]^t$ and $[1, 0]^t$. The

one-dimensional assumption implies that $S_0\tilde{v} = \lambda_0\tilde{v}$, $S_1\tilde{v} = \lambda_1\tilde{v}$ for some constants λ_0, λ_1. Then it follows that

$$\lambda_0 = c_0, \quad \lambda_1 = c_3,$$

and

$$\tilde{v} = c' \begin{bmatrix} 1 - 2c_0 - c_3 \\ c_3 \end{bmatrix} = c'' \begin{bmatrix} c_0 \\ 1 - c_0 - 2c_3 \end{bmatrix}$$

for some constants c', c''. Thus

$$(1 - 2c_0 - c_3)(1 - c_0 - 2c_3) = c_0c_3,$$

and we have either $(1 - c_0 - c_3) = 0$ or $(1 - 2c_0 - 2c_3) = 0$. Since $(1 - c_0 - c_3) \neq 0$, we must have $(1 - 2c_0 - 2c_3) = 0$, so $\tilde{v} = c'''[1, 1]^t$. By the formula of $\tilde{v}$ in (3.5), we have

$$c_0(c_0 - 1)(1 + c_0 - c_3) = c_3(1 - c_3)(1 - c_0 + c_3).$$

Simplifying this, we end up with $0 = 3/8$, which is a contradiction. □

THEOREM 3.5. *Let $1 \leq p < \infty$. Suppose that $c_0 + c_2 = c_1 + c_3 = 1$ and one of c_0, c_3, or $1 - c_0 - c_3$ is zero; then equation* (3.4) *has nonzero L^p_c-solutions if and only if*

$$|c_0|^p + |c_3|^p + |1 - c_0 - c_3|^p < 2. \tag{3.7}$$

Proof. Let $||\cdot||$ be the l^p-norm on $\mathbf{R}^2$. In view of Lemmas 3.3 and 3.4, we can assume that $\tilde{v} \neq 0$ and $H(\tilde{v})$ is one-dimensional. We first consider $c_0 = 0$. The fact that $S_0u = (1 - c_3)u$, $S_1u = c_3u$ for any $u \in H(\tilde{v})$ yields

$$\frac{1}{2^l} \sum_{|J|=l} ||S_Ju||^p = \frac{1}{2^l}(|1 - c_3|^p + |c_3|^p)^l.$$

Now apply Theorem 3.2. We see that equation (3.4) has nonzero L^p_c-solutions if and only if $|1 - c_3|^p + |c_3|^p < 2$.

Similarly, we can show that the corresponding conditions for $c_3 = 0$ and $1 - c_0 - c_3 = 0$ are $|c_0|^p + |1 - c_0|^p < 2$ and $|c_0|^p + |c_3|^p < 2$, respectively. This completes the proof. □

The following is an improvement of Theorem 3.2.

THEOREM 3.6. *For $1 \leq p < \infty$, equation* (3.4) *has a nonzero L^p_c-solution if and only if either $(c_0, c_3) = (1, 1)$ or there exists an integer l such that*

$$\frac{1}{2^l} \sum_{|J|=l} ||S_Ju||^p < 1 \quad \text{for all } u \in \mathbf{R}^2 \quad \text{and } ||u|| \leq 1. \tag{3.8}$$

Proof. By Theorem 3.2, we need only show that condition (3.8) holds when $H(\tilde{v})$ is zero or one-dimensional.

The case when $H(\tilde{v}) = \{0\}$ is obvious by Lemma 3.3, so we suppose that $H(\tilde{v})$ is one-dimensional. Equation (3.7) implies that

$$\frac{1}{2}(||S_0u||_p^p + ||S_1u||_p^p) < 1$$

for $u = [0, 1]^t$ and $[1, 0]^t$, which is a basis of $\mathbf{R}^2$; therefore the theorem follows by Remark 2. □

As a special case, we have the following corollary.

COROLLARY 3.7. *Let* $1 \le p < \infty$. *Suppose* $c_0 + c_2 = c_1 + c_3 = 1$ *and*

$$|c_0|^p + |c_3|^p + |1 - c_0 - c_3|^p < 2; \tag{3.9}$$

then (3.4) *has nonzero* L^p_c*-solutions.*

Proof. Let $||.||$ be the l^p-norm. As mentioned in Remark 2, we need only verify condition (3.6) for a basis of $\mathbf{R}^2$. Therefore, condition (3.9) implies (3.8) for $u = [1, 0]^t$ and $[0, 1]^t$ with $l = 1$. □

Similarly, we can take l to be other integers and obtain sufficient conditions for (3.4) to have nonzero L^p_c-solutions. However, the expression is more complicated. For example, for $p = 1$ the condition of (3.8) for $l = 2$ is equivalent to

$$\begin{aligned} &c_0^2 + c_3^2 + (1 - c_0 - c_3)^2 + |c_0(1 - c_0)| + |c_3(1 - c_3)| + |c_0(1 - c_0 - c_3)| \\ &\quad + |c_3(1 - c_0 - c_3)| < 4. \end{aligned} \tag{3.10}$$

In the appendix we will plot the different regions of (c_0, c_3) that admit solutions. They include the ones determined by (3.9) and (3.10), and some other known regions.

4. L^2-solutions. In this section we will show that condition (2.4) in Theorem 2.6 can be reduced to a more explicit form for the case when $p = 2$. We will use the Euclidean norm on $\mathbf{R}^N$. For any $u \in \mathbf{R}^N$, let $A_k(u) := (1/2^k) \sum_{|J|=k} ||T_J u||^2$; then

$$\begin{aligned} A_k(u) &= \frac{1}{2^k} \sum_{|J|=k} ||T_J u||^2 = \frac{1}{2^k} \sum_{|J|=k} u^t T_J^t T_J u \\ &= \frac{1}{2^k} u^t \left(\sum_{|J|=k} T_J^t T_J \right) u = \frac{1}{2^k} u^t M_k u, \end{aligned} \tag{4.1}$$

where $M_k := \sum_{|J|=k} T_J^t T_J$, and M_0 is the identity matrix. Since $T_J = T_{(J',0)}$ or $T_{(J',1)}$ for some J', it is easy to see that M_k satisfies the inductive identity

$$M_{k+1} = T_0^t M_k T_0 + T_1^t M_k T_1.$$

The matrix M_k is actually determined by its first column; its explicit form is given as follows.

LEMMA 4.1. *For any integer* $k \ge 0$, M_k *has the following form*:

$$M_k = [\alpha^{(k)}_{|i-j|}] = \begin{bmatrix} \alpha_0^{(k)} & \alpha_1^{(k)} & \cdots & \alpha_{N-1}^{(k)} \\ \alpha_1^{(k)} & \alpha_0^{(k)} & \cdots & \alpha_{N-2}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N-1}^{(k)} & \alpha_{N-2}^{(k)} & \cdots & \alpha_0^{(k)} \end{bmatrix}.$$

If we let $\alpha^{(k)} = [\alpha_0^{(k)}, \ldots, \alpha_{N-1}^{(k)}]^t$, *then* $\alpha^{(k)} = W\alpha^{(k-1)} = W^k e_1$, *where* $e_1 = [1, 0 \ldots, 0]$, *and* W *is an* $N \times N$ *matrix with*

$$(W)_{i,j} = \sum_{m=-\infty}^{\infty} c_{i+m} c_{m \pm 2j}, \quad 0 \le i, j \le N - 1,$$

where

$$c_{i\pm 2j} := \begin{cases} c_i & \text{if } j = 0, \\ c_{i+2j} + c_{i-2j} & \text{if } j \neq 0. \end{cases}$$

Proof. We prove the lemma by induction. Supposing M_k has the form as given, let $\alpha_{-l}^{(k)} = \alpha_l^{(k)}$; by using $M_{k+1} = T_0^t M_k T_0 + T_1^t M_k T_1$, the (i,j) entry of M_{k+1} is given by

$$\begin{aligned}
(M_{k+1})_{i,j} &= \sum_{m=1}^{N}\sum_{n=1}^{N} c_{2m-i-1}\alpha_{m-n}^{(k)} c_{2n-j-1} + \sum_{m=1}^{N}\sum_{n=1}^{N} c_{2m-i}\alpha_{m-n}^{(k)} c_{2n-j} \\
&= \sum_{n=1}^{N}\sum_{l=-(N-1)}^{N-1} c_{2n+2l-i-1}c_{2n-j-1}\alpha_l^{(k)} + \sum_{n=1}^{N}\sum_{l=-(N-1)}^{N-1} c_{2n+2l-i}c_{2n-j}\alpha_l^{(k)} \\
&\qquad (l = m-n) \\
&= \sum_{m=1}^{2N}\sum_{l=-(N-1)}^{N-1} c_{m+2l-i}c_{m-j}\alpha_l^{(k)} \qquad (m = 2n \text{ or } m = 2n-1) \\
&= \sum_{l=-(N-1)}^{N-1}\sum_{m=-\infty}^{\infty} c_{m+2l}c_{m-j+i}\alpha_l^{(k)} \\
&= \sum_{l=0}^{N-1}\sum_{m=-\infty}^{\infty} c_{m\pm 2l}c_{m-j+i}\alpha_l^{(k)}.
\end{aligned}$$

(We can extend the sum from $-\infty$ to $+\infty$ since c_n vanishes for $n \notin \{0, \dots N-1\}$.) By the symmetry of the range of l and m, we can rewrite the above equation as

$$(M_{k+1})_{i,j} = \sum_{l=0}^{N-1}\sum_{m=-\infty}^{\infty} c_{m\pm 2l}c_{m-i+j}\alpha_l^{(k)}.$$

Hence $(M_{k+1})_{i,j} = (M_{k+1})_{i+1,j+1} = (M_{k+1})_{j,i}$, and M_k has the form as asserted. Also from the proof above, we see that

$$\alpha_i^{(k+1)} = (M_{k+1})_{i,0} = \sum_{l=0}^{N-1}\sum_{m=-\infty}^{\infty} c_{m\pm 2l}c_{m-i}\alpha_l^{(k)}.$$

Therefore we may write

$$\begin{bmatrix} \alpha_0^{(k+1)} \\ \alpha_1^{(k+1)} \\ \vdots \\ \alpha_{N-1}^{(k+1)} \end{bmatrix} = W \begin{bmatrix} \alpha_0^{(k)} \\ \alpha_1^{(k+1)} \\ \vdots \\ \alpha_{N-1}^{(k)} \end{bmatrix} = W^{k+1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where W is an $N \times N$ matrix with (i,j) entry as

$$(W)_{i,j} = \sum_{m=-\infty}^{\infty} c_{i+m}c_{m\pm 2j}, \quad 0 \le i,j \le N-1. \qquad \square$$

PROPOSITION 4.2. *The matrix W can be written as the product $A \cdot B$, where*

$$A = [c_{i+j}]_{\substack{0\le i\le N-1,\\ |j|\le N-1}}, \quad B = [c_{i\pm 2j}]_{\substack{|i|\le N-1,\\ 0\le j\le N-1}}.$$

For any vectors $u = [u_0, \dots, u_{N-1}]^t$, let $\Psi(u)$ be the vector

$$\Psi(u) = \left[\sum u_i^2, 2\sum u_i u_{i+1}, \dots, 2\sum u_i u_{i+N-1}\right]^t;$$

then Theorem 2.6 can be written as follows.

PROPOSITION 4.3. *Equation (2.1) has a nonzero L^2_c-solution if and only if there is a 2-eigenvector v of $(T_0 + T_1)$ such that for any $u \in H(\tilde{v})$,*

$$\lim_{l\longrightarrow\infty} \frac{1}{2^l}\Psi(u)^t \cdot W^l \cdot \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0. \tag{4.2}$$

Proof. For any $u = [u_0, \dots, u_{N-1}]^t \in H(\tilde{v})$, by (4.1) and Lemma 4.1, we have

$$\begin{aligned} A_l(u) &= \frac{1}{2^l}\sum_{|J|=l} ||T_J u||^2 = \frac{1}{2^l} u^t M_l u \\ &= \frac{1}{2^l}[u_0, \dots, u_{N-1}] M_l \begin{bmatrix} u_0 \\ \vdots \\ u_{N-1} \end{bmatrix} = \frac{1}{2^l}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} u_i u_j \alpha^{(l)}_{|i-j|} \\ &= \frac{1}{2^l}\Psi(u)^t \cdot \begin{bmatrix} \alpha_0^{(l)} \\ \vdots \\ \alpha_{N-1}^{(l)} \end{bmatrix} = \frac{1}{2^l}\Psi(u)^t \cdot W^l \cdot \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \end{aligned}$$

Now, apply Theorem 2.6 and the proof is complete. □

We now assume $\sum c_{2n} = \sum c_{2n+1} = 1$; then the vector $[1, \dots, 1]^t$ is a right eigenvector of W with eigenvalue 2. Indeed, for any $0 \le i \le N-1$, the sum of the ith row equals

$$\begin{aligned} \sum_{j=0}^{N-1} W_{i,j} &= \sum_{j=0}^{N-1}\sum_{m=-N}^{N} c_{i+m}c_{m\pm 2j} = \sum_{j=-(N-1)}^{N-1}\sum_{m=-N}^{N} c_{i+m}c_{m+2j} \\ &= \sum_{m=-N}^{N} c_{i+m}\left(\sum_{j=-(N-1)}^{N-1} c_{m+2j}\right) = \sum_{m=-N}^{N} c_{i+m} = 2. \end{aligned}$$

Also, $[1, \dots, 1]W^t u = [2, \dots, 2]u = 2\sum u_i$ implies that W^t is invariant on H.

Recall that the *algebraic multiplicity* of an eigenvalue λ_0 is the order of the factor $(\lambda - \lambda_0)$ in the characteristic polynomial. We can now state and prove our main theorem of this section.

THEOREM 4.4. *Suppose $\sum c_{2m} = \sum c_{2m+1} = 1$. If the eigenvalue 2 of W is of algebraic multiplicity 1, and all other eigenvalues of W are less than 2 in absolute value, then equation (2.1) has nonzero L^2_c-solutions.*

Proof. For any $u = [u_0, \dots, u_{N-1}]^t \in H(\tilde{v}) \subseteq H$,

$$\sum_k \sum_{|i-j|=k} u_i u_j = \left|\sum u_j\right|^2 = 0,$$

so $\Psi(u) \in H$. It follows from the assumption that the eigenvalues of W on H are less than 2 that we have for $w \in H$,

$$\lim_{l\to\infty} \frac{1}{2^l} w^t \cdot W^l = 0.$$

Theorem 4.4 now follows from Proposition 4.3 directly. □

Remark 7. We can write the matrix W as follows: Let $P = [p_0, \dots, p_{N-1}]$ be an orthonormal matrix with $p_0 = [\frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}}]^t$; it follows that

$$P^* W P = \begin{bmatrix} 2 & * \\ 0 & W_1 \end{bmatrix},$$

where the $(N-1) \times (N-1)$ matrix W_1 is the restriction of W on H. Theorem 4.4 tells us that equation (2.1) has a nonzero L^2_c-solution if W_1 has spectral radius less than 2.

For the converse of the above theorem we need the following lemma.

LEMMA 4.5. *The image of H under the map Ψ contains an $(N-1)$-dimensional region of H.*

Proof. This follows from the observation that the vectors

$$[2, -2, 0, \dots, 0]^t,\ [2, 0, -2, 0, \dots, 0]^t,\ \dots,\ [2, 0, \dots, 0, -2]^t$$

are the images of

$$[1, -1, 0, \dots, 0]^t,\ [1, 0, -1, 0, \dots, 0]^t,\ \dots,\ [1, 0, \dots, 0, -1]^t$$

under the continuous map Ψ. □

PROPOSITION 4.6. *Suppose $\sum c_{2n} = \sum c_{2n+1} = 1$ and (2.1) has a nonzero L^2_c-solution f. Let $v = [f_{[0,1]}, \dots, f_{[0,N-1]}]^t$ be the average vector of f; if $H(\tilde{v}) = H$ and $\{W^k e_1\}_{k=1}^N$ spans $\mathbf{R}^N$, then the eigenvalue 2 of W has algebraic multiplicity 1, and all other eigenvalues are less than 2 in absolute value.*

Proof. If $\{W^k e_1\}$ spans $\mathbf{R}^N$, then (4.2) is equivalent to

$$\lim_{l\to\infty} \frac{1}{2^l} \Psi(u)^t \cdot W^l = 0$$

for any $u \in H(\tilde{v})$. But if $H(\tilde{v}) = H$, then by Lemma 4.5, $\Psi(H)$ is also a $(N-1)$-dimensional region contained in H. So the spectral radius of W_1 must be less than 2 and the proposition follows. □

Remark 8. If we impose the m-sum rules (2.7), then for any $u \in H_m$, we also have $\Psi(u) \in H_m$. This is true because for any $j = 0, 1, \dots, m$,

$$\Psi(u)^t \cdot \begin{bmatrix} 0^j \\ 1^j \\ \vdots \\ (N-1)^j \end{bmatrix} = \sum_{k=0}^{N-1} k^j \left(\sum_i u_i u_{i+k} + \sum_i u_i u_{i-k} \right)$$

$$= \sum_i u_i \left(\sum_{k=0}^{N-1} k^j u_{i+k} \right) + \sum_i u_i \left(\sum_{k=0}^{N-1} k^j u_{i-k} \right),$$

which is 0 since it can be shown inductively that $\sum_{k=0}^{N-1} k^j u_{i\pm k} = 0$ for $j = 0, 1, \ldots, m$ and any i. By Theorem 2.12, Proposition 4.3 reduces to the following corollary.

COROLLARY 4.7. *Suppose that* $\sum c_n = 2$ *and the m-sum rules hold, and suppose there is a 2-eigenvector* v *of* $(T_0 + T_1)$ *such that for any* $u \in H_m$,

$$\lim_{l \longrightarrow \infty} \frac{1}{2^l} u^t \cdot W^l \cdot \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0.$$

Then equation (2.1) *has a nonzero* L^2_c*-solution.*

Remark 9. In [E], [Her1], [Her2], and [V] there are various characterizations of the existence of L^2_c-solutions; the Sobolev exponents and energy moments are also obtained. In those papers the Fourier method was used; the dilation equation (2.1) becomes

$$\hat{f}(\xi) = m_0\left(\frac{\xi}{2}\right)\hat{f}\left(\frac{\xi}{2}\right),$$

where $m_0(\xi) = \frac{1}{2}\sum c_k e^{-ik\xi}$. Let $g(\xi) = \sum_{k=-\infty}^{\infty} |\hat{f}(\xi + 2\pi k)|^2$; then

$$g(\xi) = \left|m_0\left(\frac{\xi}{2}\right)\right|^2 g\left(\frac{\xi}{2}\right) + \left|m_0\left(\frac{\xi}{2} + \frac{\pi}{2}\right)\right|^2 g\left(\frac{\xi}{2} + \frac{\pi}{2}\right). \tag{4.3}$$

Villemoes [V] showed that a nonzero L^2_c-solution exists if and only if there is a nonnegative trigonometric polynomial $g(\xi) = \sum_{k=-(N-1)}^{N-1} a_k e^{-ik\xi}$ satisfying $g(0) > 0$ and (4.3).

For $g(\xi) = \sum_{k=-(N-1)}^{N-1} a_k e^{-ik\xi}$, it follows from a direct calculation that equation (4.3) is equivalent to the fact that $[a_{-(N-1)}, \ldots, a_{N-1}]$ is a left 2-eigenvector of W', where W' is a $(2N-1) \times (2N-1)$ matrix with (i, j) entry equal to

$$\sum_{-\infty}^{\infty} c_{i+m} c_{m+2j}, \quad -(N-1) \le i, j \le N-1.$$

By the symmetry of W' and the fact that the c_k's are real, one can reduce the operator W' to the matrix W we consider here (see Remark 3.2 in [V]).

Lawton [La] showed that the scaling function f generates an orthonormal basis of L^2 if and only if the vector $[a_{-(N-1)}, \ldots, a_{N-1}]$ with $a_k = \delta_{0,k}$ is the only left eigenvector of W' corresponding to eigenvalue 2.

Hervé [Her1] and [Her2] used an iteration argument based on (4.3) (with 2 replaced by p) to determine the condition for the existence of the solution whose Fourier transform $\hat{f}$ is in L^p. He also calculated the Sobolev exponents

$$s_p = \sup\left\{s \ge 0 : \quad \int |\hat{f}(\xi)|^p (1 + |\xi|^{ps}) d\xi < \infty\right\}$$

for such f.

To conclude this section we will demonstrate the foregoing results for the case $N = 3$. By Lemma 4.1, we calculate that

$$W = \begin{bmatrix} 2(1-\delta) & 2\delta & 0 \\ 1 - c_0 c_3 & 1 & c_0 c_3 \\ \delta & 2(1-\delta) & \delta \end{bmatrix},$$

where $\delta = c_0 - c_0^2 + c_3 - c_3^2$. In addition to 2, W has two eigenvalues

$$\frac{1}{2}\left(1 - c_0 + c_0^2 - c_3 + c_3^2 \pm \sqrt{1 + 26c_0c_3 - (c_0 + c_3)\omega_1 + (c_0^2 + c_3^2)\omega_2}\right),$$

where

$$\omega_1 = (18c_0^2 + 18c_3^2 - 16c_0c_3 + 6), \quad \omega_2 = (9c_0^2 + 9c_3^2 + 16c_0c_3 + 15).$$

It follows from Theorem 4.4 that if the two eigenvalues are less then 2, then L_c^2-solutions exist.

For $N = 3$, if we adopt the approach in §3 by reducing the matrices T_i on H to S_i on $\mathbf{R}^2$, $i = 0, 1$, then the above analysis is more transparent and the result can be sharpened.

Let $\tilde{M}_0$ be the 2×2 identity matrix. Assume

$$\tilde{M}_k = \sum_{|J|=k} S_J^t S_J = \begin{bmatrix} \alpha^{(k)} & \beta^{(k)} \\ \beta^{(k)} & \alpha^{(k)} \end{bmatrix}.$$

A direct computation shows that

$$\begin{aligned} \tilde{M}_{k+1} &:= \sum_{|J|=k+1} S_J^t S_J = S_0^t \tilde{M}_k S_0 + S_1^t \tilde{M}_k S_1 \\ &= \begin{bmatrix} (c_0^2 + c_3^2 + d^2)\alpha^{(k)} - 2c_0c_3\beta^{(k)} & -(c_0d + c_3d)\alpha^{(k)} + (c_0d + c_3d)\beta^{(k)} \\ -(c_0d + c_3d)\alpha^{(k)} + (c_0d + c_3d)\beta^{(k)} & (c_0^2 + c_3^2 + d^2)\alpha^{(k)} - 2c_0c_3\beta^{(k)} \end{bmatrix}, \end{aligned}$$

where $d = (1 - c_0 - c_3)$. Comparing the first columns of the two matrices $\tilde{M}_k$ and $\tilde{M}_{k+1}$, we can define the matrix $\tilde{W}$ as follows:

$$\begin{bmatrix} \alpha^{(k+1)} \\ \beta^{(k+1)} \end{bmatrix} = \tilde{W} \begin{bmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{bmatrix} = \tilde{W}^{k+1} \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

where

$$\tilde{W} = \begin{bmatrix} c_0^2 + c_3^2 + d^2 & -2c_0c_3 \\ -d(c_0 + c_3) & d(c_0 + c_3) \end{bmatrix}.$$

A direct computation shows that $\tilde{W}$ has the same eigenvalues as W_1 in Remark 7; however, we do not known their exact relationship.

THEOREM 4.8. *Suppose $c_0 + c_2 = c_1 + c_3 = 1$; then the dilation equation (2.1) with $N = 3$ has nonzero L_c^2-solutions if and only if either $(c_0, c_3) = (1, 1)$ or the matrix*

$$\begin{bmatrix} c_0^2 + c_3^2 + d^2 & -2c_0c_3 \\ -d(c_0 + c_3) & d(c_0 + c_3) \end{bmatrix},$$

where $d = (1 - c_0 - c_3)$, has spectral radius less than 2.

Proof. For any $u = [x, y]^t$, we have

$$\begin{aligned} \frac{1}{2^l}\sum_{|J|=l} ||S_J u||^2 &= \frac{1}{2^l}\sum_{|J|=l} u^t S_J^t S_J u = \frac{1}{2^l} u^t \left(\sum_{|J|=l} S_J^t S_J\right) u \\ &= \frac{1}{2^l} u^t \tilde{M}_l u = \frac{1}{2^l}[x, y] \begin{bmatrix} \alpha^{(l)} & \beta^{(l)} \\ \beta^{(l)} & \alpha^{(l)} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= \frac{1}{2^l}[x^2 + y^2, 2xy] \begin{bmatrix} \alpha^{(l)} \\ \beta^{(l)} \end{bmatrix} \\ &= \frac{1}{2^l}[x^2 + y^2, 2xy]\tilde{W}^l \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \end{aligned}$$

Since $\{[x^2+y^2, 2xy];\quad [x,y]\in \mathbf{R}^2\}$ spans $\mathbf{R}^2$, the condition

$$\frac{1}{2^l}\sum_{|J|=l}||S_J u||^2 \longrightarrow 0 \quad \text{as} \quad l\longrightarrow\infty \quad \text{for all} \quad u\in\mathbf{R}^2$$

is equivalent to

$$\frac{1}{2^l}\tilde{W}^l\begin{bmatrix}1\\0\end{bmatrix}\longrightarrow 0 \quad \text{as} \quad l\longrightarrow\infty. \tag{4.4}$$

We will show that (4.4) holds if and only if the spectral radius of $\tilde{W}$ is less than 2; hence Theorem 3.6 applies and we are done.

If $d(c_0+c_3)\neq 0$, let $u_1=\tilde{W}[1,0]^t$; then u_1 and $[1,0]^t$ are linearly independent. (4.4) is equivalent to the statement that $\frac{1}{2^l}\tilde{W}^l u\longrightarrow 0$ for all $u\in\mathbf{R}^2$, and hence $\frac{1}{2^l}\tilde{W}^l\longrightarrow 0$ as $l\longrightarrow\infty$. This means that the eigenvalues of W are less than 2.

If $d(c_0+c_3)=0$, then $\tilde{W}$ is of the form

$$\begin{bmatrix}w_1 & w_2\\ 0 & 0\end{bmatrix}$$

for some w_1 and w_2. It follows that

$$\tilde{W}^l=\begin{bmatrix}w_1^l & w_1^{l-1}w_2\\ 0 & 0\end{bmatrix},$$

and (4.4) implies $|(1/2^l)w_1^l|<1$, that is, $|w_1|<2$. Again the eigenvalues of W are all less than 2 in absolute value. □

Appendix. For the four-coefficient dilation equation

$$f(x)=c_0f(2x)+c_1f(2x-1)+c_2f(2x-2)+c_3f(2x-3) \tag{A1}$$

with $c_0+c_2=1$, $c_1+c_3=1$, let c_0, c_3 be the independent parameters. We use the **Mathematica** on a NeXT workstation to plot the following regions of (c_0,c_3), for which the compactly supported L^1 and L^2 solutions exist.

Let D_l be the regions of (c_0,c_3) for which

$$\frac{1}{2^l}\sum_{|J|=l}||S_J u||<1 \quad \text{for} \quad u\in\mathbf{R}^2,\ ||u||\le 1 \tag{A2}$$

holds. By Theorem 3.6, except for $(c_0,c_3)=(1,1)$, equation (A1) has a nonzero compactly supported L^1-solution if and only if (c_0,c_3) is in the union of the regions D_l, $l=1,2,\ldots$. In Fig. 1 we display the regions D_l for $l=1,2,4$, and 8. Here the norm is $||[x,y]^t||=|x|+|y|$.

Note that the regions are increasing (Remark 5, Theorem 2.6). When $l=1$ and 2, condition (A2) can be written as

$$|c_0|+|c_3|+|1-c_0-c_3|<2,$$

and

$$c_0^2+c_3^2+(1-c_0-c_3)^2+|c_0(1-c_0)|+|c_3(1-c_3)|+|c_0(1-c_0-c_3)|$$
$$+|c_3(1-c_0-c_3)|<4,$$

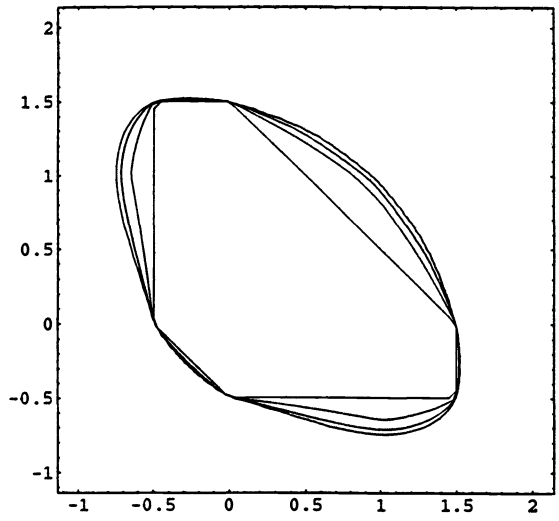


FIG. 1.

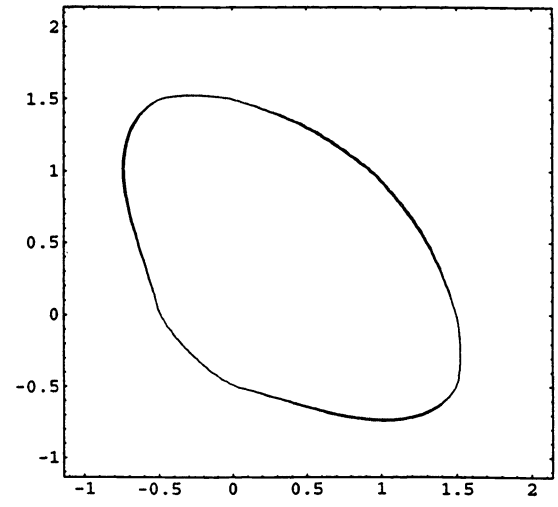


FIG. 2.

respectively. For $l \geq 3$, the expression is more tedious.

In Fig. 2 we plot the regions D_l for $l = 6$ and 8. Note that they are very close, and hence they are good approximations of the admissible region of (c_0, c_3) for L^1-solutions.

In Fig. 3 we plot the following regions of (c_0, c_3) for the existence of the L^1-solutions from some previous results.

The region outside the ellipse

$$c_0^2 + c_3^2 - c_0 - c_3 + c_0 c_3 = 1$$

is known to have no L^1-solution for (A1).

The region bounded by the dotted line

$$c_0^2 + c_3^2 + |c_0(1 - c_0)| + |c_3(1 - c_3)| + 2|1 - c_0 - c_3| < 4$$

is a sufficient condition given by Pan [P].

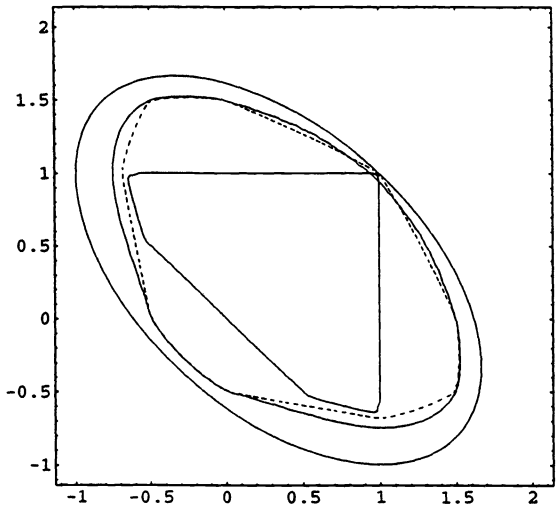


FIG. 3.

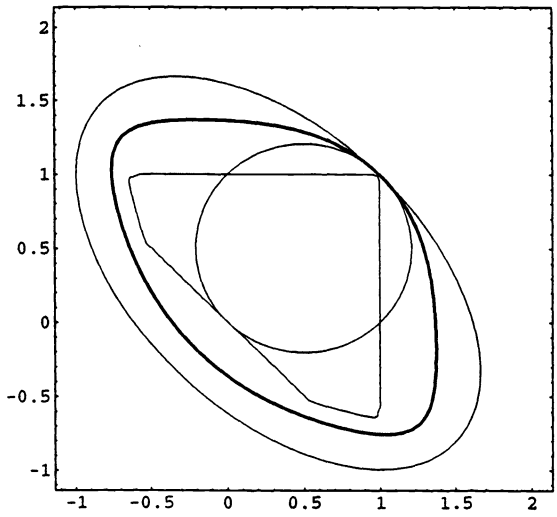


FIG. 4.

The region D_8 is determined by (A2) with $l = 8$.

Also, the triangular-shaped region approximates the domain where the joint spectral radius of T_0 and T_1 is less than 1, hence nonzero compactly supported continuous solutions exist there.

In Fig. 4, we plot the following regions:

First we plot the region determined by the ellipse as in Fig. 3.

Next we plot the region bounded by the thicker line consisting of points (c_0, c_3) for which the matrix

$$\begin{bmatrix} c_0^2 + c_3^2 + d^2 & -2c_0c_3 \\ -d(c_0 + c_3) & d(c_0 + c_3) \end{bmatrix}, \quad \text{where} \quad d = 1 - c_0 - c_3,$$

has spectral radius less than 2. This is a necessary and sufficient condition for (A1) to have nonzero compactly supported L^2-solutions with one exception: $(c_0, c_3) = (1, 1)$ (Theorem 4.8).

We also plot the region for the existence of compactly supported continuous solutions as in Fig. 3.

Finally, we plot the circular region

$$(c_0 - 1/2)^2 + (c_3 - 1/2)^2 \leq 1/2,$$

a sufficient condition of the existence of L^2-solutions given in [La]. The boundary is called the *circle of orthogonality*: if the wavelet generated by the scaling function satisfying (A1) is orthonormal, then the point (c_0, c_3) must be on the circle.

Acknowledgment. The authors express their gratitude to the participants of the wavelet seminar for many valuable discussions, especially to Professor Y. Pan who shared his idea which motivated our paper. The authors also thank the referees for many valuable comments and for providing the relevant literature for the L^2-case.

REFERENCES

[BW1] M. A. BERGER AND Y. WANG, *Two-scale dilation equations and cascade algorithm*, Random Comput. Dynamics, to appear.

[CH1] D. COLELLA AND C. HEIL, *The characterization of continuous, four-coefficient scaling functions and wavelets*, IEEE. Trans. Inform. Theory, 30 (1992), pp. 876–881.

[CH2] ———, *Characterizations of scaling functions, I. Continuous solutions*, J. Math. Anal. Appl., 15 (1994), pp. 496–518.

[D] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[DL1] I. DAUBECHIES AND J. LAGARIAS, *Two-scale difference equation I. Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[DL2] ———, *Two-scale difference equation II. Local regularity, infinite products of matrices, and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[dR] G. DE RHAM, *Sur un example de fonction continue sans dérivée*, Enseign. Math., 3 (1957), pp. 71–72.

[E] T. EIROLA, *Sobolev characterization of solutions of dilation equadtions*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[H] C. HEIL, *Methods of solving dilation equations*, Proc. 1991 NATO Adv. Sci. Ins. on Prob. and Stoch. Methods in Anal. with Appl., J. Byrnes, ed., Kluwer Academic Publishers, Dordrecht, 1993.

[Her1] L. HERVÉ, *Méthodes d'opérateurs quasi-compacts en analyse multirésolution, applications à la construction d'ondelettes et à l'interpolation*, Thèse, Laboratoire de Probabilités, Université de Rennes I, 1992.

[Her2] ———, *Construction et régularité des fonctions d'échelle*, Laboratoire de Probabilités, Université de Rennes I, preprint.

[Hu] J. HUTCHINSON, *Fractals and self-similarity*, Indiana Univ. Math. J., 30 (1981), pp. 713–747.

[JRS] P. JANARDHAN, D. ROSENBLUM, AND R. STRICHARTZ, *Numerical experiments in Fourier asymptotics of Cantor measures and wavelets*, Experiment. Math., 1 (1992), pp. 249–273.

[L1] K. LAU, *Fractal measures and mean p-variations*, J. Funct. Anal., 108 (1992), pp. 427–457.

[L2] ———, *Dimension of a family of singular Bernoulli convolutions*, J. Funct. Anal., 116 (1993), pp. 335–358.

[La] W. LAWTON, *Necessary and sufficient conditions for constructing orthonormal wavelet bases*, J. Math. Phys., 32 (1991), pp. 57–64.

[M] S. MALLAT, *Multiresolution approximation and wavelet orthonormal bases for $L^2(R)$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[MP] C. A. MICHELLI AND H. PRAUTZSCH, *Uniform refinement of curves*, Linear Algebra Appl., 114/115 (1989), pp. 841–870.

[P] Y. PAN, *On the existence of L^1 scaling functions*, J. Math. Phy. Sci., to appear.

[RS] G. ROTA AND G. STRANG, *A note on the joint spectral radius*, Indag. Math., 22 (1960), pp. 379–381.

[R] H. L. ROYDEN, *Real Analysis*, 3rd edition, Macmillan, New York, 1988.

[S] G. STRANG, *Wavelets and dilation equations: A brief introduction*, SIAM Rev., 31 (1989), pp. 614–627.

[Str] R. STRICHARTZ, *Self-similar measures and their Fourier transforms II*, Trans. Amer. Math. Soc., 336 (1993), pp. 336–361.

[V] L. F. VILLEMOES, *Energy moments in times and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.

[W] Y. WANG, *Two-scale dilation equations and mean spectral radius*, Random Comput. Dynamics, to appear.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 1047–1060, July 1995

012

INTERVAL OSCILLATION CONDITIONS FOR DIFFERENCE EQUATIONS*

Q. KONG† AND A. ZETTL†

Abstract. Interval-type sufficient conditions for oscillation of solutions of second-order difference equations are established. These conditions are new even though their analogues for differential equations have been known for some time.

Key words. oscillation, difference equations, linear, second order

AMS subject classifications. 39A10, 39A12

1. Introduction.

We study oscillatory properties of solutions of the symmetric second-order linear difference equation

$$\Delta(p_n\Delta x_n) + q_n x_{n+1} = 0, \quad n = 1, 2, \ldots, \tag{1.1}$$

where Δ is the forward difference operator $\Delta x_n = x_{n+1} - x_n$, and $p = \{p_n;\ n \in \mathbb{N}\}$ and $q = \{q_n;\ n \in \mathbb{N}\}$ are sequences of real numbers with $p_n > 0$ for $n \in \mathbb{N} = \{1, 2, \ldots\}$.

The oscillation problem for equation (1.1) has been considered by many authors [1]–[9], [11]–[14], some for the equivalent form

$$p_n x_{n+1} + p_{n-1}x_{n-1} = b_n x_n, \ n = 1, 2, \ldots, \tag{1.2}$$

where $b_n = p_n + p_{n-1} - q_{n-1}$.

In view of the extensive literature concerning the corresponding second-order scalar differential equation

$$(p(t)y'(t))' + q(t)y(t) = 0, \tag{1.3}$$

it is interesting to obtain discrete analogues of known oscillation criteria and note the similarities and differences which arise between the continuous and discrete cases.

The main purpose of this paper is to establish a discrete analogue of the telescoping principle of Kwong and Zettl [10] for equation (1.3), which is surprisingly useful for establishing and improving oscillation criteria, and apply it to obtain some new results on the oscillation of equation (1.1).

DEFINITION 1.1. *A sequence $x = \{x_n;\ n \in \mathbb{N}\}$ is said to be nonoscillatory if there exists an integer n_1 such that for all $n \geq n_1$, we have that $x_n x_{n+1} > 0$. Otherwise, the sequence x is called oscillatory.*

Since either all solutions of equation (1.1) are oscillatory or none are oscillatory (cf. [1]), equation (1.1) may be classified as oscillatory or nonoscillatory.

Our approach to the oscillation problem of equation (1.1) is based largely on a discrete version of the Riccati equation. If $x = \{x_n;\ n \in \mathbb{N}\}$ is a solution of equation (1.1) with $x_n x_{n+1} > 0$ for $n \geq n_1 \geq 1$, we let

$$u_n = -p_n\Delta x_n / x_n, \ n \geq n_1. \tag{1.4}$$

Then, since $-u_n + p_n = p_n x_{n+1}/x_n > 0$, we have

$$\Delta u_n = \frac{u_n^2}{-u_n + p_n} + q_n \tag{1.5}$$

* Received by the editors June 29, 1993; accepted for publication October 11, 1993.
† Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

or, equivalently,

$$u_{n+1} = \frac{p_n u_n}{-u_n + p_n} + q_n. \tag{1.6}$$

The cardinality of a set A is denoted by $\mathrm{card}(A)$. Denote by S the set of all real sequences $x = \{x_n;\ n \in \mathbb{N}\}$. The notation $[n, m]$ denotes the interval of $\mathbb{N}$ consisting of $\{n, n+1, \ldots, m\}$. Similarly, $(n, m]$ denotes the half-open interval including m but not n, etc.

In the following we assume

$$(H) \qquad J = \bigcup_{i=1}^{j} J_i, \qquad J_i = (a_i, b_i],\ i = 1, \ldots, j,\ j \le +\infty, \tag{1.7}$$

where $a_i,\ b_i \in \mathbb{N},\ i = 1, \ldots, j$, satisfy $a_i < b_i < a_{i+1}$ and $\mathrm{card}(\mathbb{N} \setminus J) = \infty$.

Based on the above set J we define an interval shrinking transformation $t = t_J : \mathbb{N} \to \mathbb{N}$ as follows:

$$N = t(n) = \mathrm{card}\,([1, n] \cap J^c), \tag{1.8}$$

where $J^c = \mathbb{N} \setminus J$. Let $A_i = t(a_i)$; then $A_i = t(n)$ for $n \in [a_i, b_i], i = 1, \ldots, j$. This transformation t induces a transformation $T = T_J : S \to S$ defined as follows: for $x \in S$, $x = \{x_n;\ n \in \mathbb{N}\}$,

$$Tx = X = \{X_N;\ N \in \mathbb{N}\} \text{ with } X_N = x_n \text{when } t(n) = N. \tag{1.9}$$

2. Telescoping principle. The following lemma is from Chen and Erbe [1].

LEMMA 2.1. *A sequence $x \in S$ is a solution of equation* (1.1) *satisfying $x_n x_{n+1} > 0$ for $n_1 \le n \le n_2 \le \infty$ if and only if the corresponding solution u of equations* (1.5) *or* (1.6) *defined by* (1.4) *satisfies $u_n < p_n$, $n_1 \le n \le n_2$.*

Our first result is a type of comparison theorem.

THEOREM 2.2. *Let $p_n > 0$ for $n \in \mathbb{N}$, and assume (H) holds. Let $P = Tp$, $Q = Tq$ for $T = T_J$. Assume*

$$\sum_{n=a_i+1}^{b_i} q_n \ge 0,\ i = 1, \ldots, j. \tag{2.1}$$

Suppose $Y = \{Y_N;\ N \in \mathbb{N}\}$ is a solution of the equation

$$\Delta(P_N \Delta Y_N) + Q_N Y_{N+1} = 0,\ N = 1, 2, 3, \ldots, \tag{2.2}$$

such that $Y_N Y_{N+1} > 0$ for $N < K$ and $Y_K Y_{K+1} \le 0$. If $x = \{x_n\}$ is a solution of equation (1.1) *such that $x_1 \ne 0$ and $p_1 \Delta x_1 / x_1 \le P_1 \Delta Y_1 / Y_1$, then there exists $h \le k$ such that $x_h x_{h+1} \le 0$, where $K = t(k)$. More precisely, if $K \le A_i$, then there exists $h \le a_i$ such that $x_h x_{h+1} \le 0$, $i = 1, 2, \ldots, j$.*

Proof. In this proof, by $x \not\asymp Y$ we mean either $x \ge Y$ or x does not exist. The proof is by induction. Assume the conclusion is not true. Then $u = \{u_n\}$ defined by (1.4) satisfies (1.5) and (1.6) for $n = 1, \ldots, k$ and $u_n < p_n$, $n = 1, \ldots, k$. Let $V_N = -P_N \Delta Y_N / Y_N$. Then

$$V_{N+1} = \frac{P_N V_N}{-V_N + P_N} + Q_N,\ N = 1, \ldots, K-1, \tag{2.3}$$

and by Lemma 2.1, $V_N < P_N$, $N = 1, \dots, K-1$, and $V_K \not< P_K$.

If $K \le A_1 = a_1$, then for $n = 1, \dots, K$, $N = n$, and hence $P_N = p_n$, $Q_N = q_n$, and equation (2.3) is the same as (1.6). By the hypothesis $u_1 \ge V_1$, comparing (1.6) and (2.3) step by step, we find that $u_{n+1} \ge V_{n+1}$, $n = 1, \dots, K-1$. In particular,

$$u_k = u_K \ge V_K \not< P_K = p_k.$$

This implies that $u_k \not< p_k$, contradicting the assumption.

If $A_1 < K \le A_2$, then arguing as above we see that $u_{a_1+1} = u_{A_1+1} \ge V_{A_1+1}$. Adding (1.5) for n from a_1 to b_1 and using (2.1) we obtain

$$u_{b_1+1} - u_{a_1+1} = \sum_{n=a_1+1}^{b_1} \frac{u_n^2}{-u_n + p_n} + \sum_{n=a_1+1}^{b_1} q_n \ge 0,$$

hence $u_{b_1+1} \ge u_{a_1+1} \ge V_{A_1+1}$. Noting that $t(b_1+1) = B_1 + 1$, we see that u_n, V_N satisfy the same Riccati equation for $b_1+1 \le n \le k$ and $A_1+1 \le N \le K$, respectively. As before, we see that $u_k \ge V_K \not< P_K = p_k$ and, again, this implies that $u_k \not< p_k$, contradicting the assumption. The proof of the inductive step from i to $i+1$ is similar and hence is omitted. □

THEOREM 2.3 (telescoping principle). *Under the conditions and with the notation of Theorem* 2.2, *if equation* (2.2) *is oscillatory, then equation* (1.1) *is oscillatory.*

Proof. Let $Y = \{Y_n\}$ be a solution of (2.2) with $Y_1 \ne 0$. Let $x = \{x_n\}$ be a solution of (1.1) satisfying $x_1 \ne 0$, $p_1 \Delta x_1 / x_1 \le P_1 \Delta Y_1 / Y_1$. By Theorem 2.2, there exists $h_1 > 0$ such that $x_{h_1} x_{h_1+1} \le 0$. Now, working on the solution for $n \ge h_1 + 1$ instead of $n \ge 1$ and proceeding as before, we show that there exists $h_2 \ge h_1 + 1$ such that $x_{h_2} x_{h_2+1} \le 0$. Continuing this process leads to the conclusion that x is oscillatory, hence (1.1) is oscillatory. □

This principle can be applied to get many new examples of oscillatory equations. We use a process that is the reverse of the construction in Theorem 2.2. Start with any known oscillatory equation (2.2). Choose a sequence of integers $A_i \to \infty$. Cut the plane at each vertical line $n = A_i$ and pull the two half-planes apart to form a gap of arbitrary finite length. Now fill the gap with an arbitrary positive p_i and any q_i whose sum over the length of the gap is nonnegative. Do this at each A_i and denote the new coefficient sequences thus constructed by p, q. Then equation (1.1) is oscillatory.

The telescoping principle is also useful in extending various known oscillation criteria. It implies that any sufficient conditions for oscillation need only be verified on intervals, for example, on $\mathbb{N} \setminus J$, where J is defined by (1.7); on J, p_n and q_n can be arbitrary as long as $p_n > 0$ and the q_n have a nonnegative sum over each interval $(a_i, b_i]$ of J, $i = 1, \dots, j$.

3. Perturbation of nonoscillatory equations. The telescoping principle obtained in §2 is not as convenient to use as the one for differential equations. The difficulties are caused by the jumping behavior of the sums

$$r_n = \sum_{i=1}^{n} q_i. \tag{3.1}$$

In general, the sum

$$R_N = \sum_{i=1}^{N} Q_i \tag{3.2}$$

is different from r_n and this causes considerable difficulty in applying the telescoping principle. To overcome this difficulty we introduce a technique based on a perturbation which leaves the nonoscillatory property of equation (1.1) invariant.

Equation (1.6) (or (1.5)) is said to be nonoscillatory if it has a solution $u \in S$ satisfying $u_n < p_n$, $n > n_1$ for some n_1. Assume (1.6) is nonoscillatory. The results of this section will show that by adjusting the values of p_n and q_n at the boundary points of J, the adjusted equation is also nonoscillatory and the sums r_n do not change on $\mathbb{N} \setminus J$.

LEMMA 3.1. *Assume equation* (1.6) *is nonoscillatory and* $u \in S$ *is a solution of* (1.6) *such that* $u_n < p_n$, $n \geq n_1$. *Let* $m \in \mathbb{N}$, $m \geq n_1$, *and let* c *be any real constant. Consider the equation*

$$w_{n+1} = \frac{\hat{p}_n w_n}{-w_n + \hat{p}_n} + \hat{q}_n, \quad n = 1, 2, \ldots, \tag{3.3}$$

where $\hat{p}_n = p_n$, $\hat{q}_n = q_n$ *for* $n < m$; $\hat{p}_n = p_{n-1}$, $\hat{q}_n = q_{n-1}$ *for* $n > m+1$; *and* $\hat{q}_m = q_m + c$, $\hat{q}_{m+1} = -c$. *Then, if* $\hat{p}_{m+1}$ *is sufficiently large, there exists* $\hat{p}_m > p_m$ *such that equation* (3.1) *is nonoscillatory. Furthermore,*

$$\hat{p}_m \to p_m \quad \text{as} \quad \hat{p}_{m+1} \to \infty.$$

Proof. Let $w_n = u_n$ for $n \leq m$ and $w_n = u_{n-1}$ for $n > m+1$. Then we only need to pick a $\hat{p}_m$ corresponding to $\hat{p}_{m+1}$ such that the solution of the adjusted equation satisfies

$$u_m = w_m < \hat{p}_m \text{ and } w_{m+1} < \hat{p}_{m+1}.$$

To this end, compare

$$u_{m+1} = \frac{p_m u_m}{-u_m + p_m} + q_m \tag{3.4}$$

with

$$\begin{cases} w_{m+1} = \dfrac{\hat{p}_m u_m}{-u_m + \hat{p}_m} + q_m + c, \\ u_{m+1} = w_{m+2} = \dfrac{\hat{p}_{m+1} w_{m+1}}{-w_{m+1} + \hat{p}_{m+1}} - c. \end{cases} \tag{3.5}$$

For a sufficiently large $\hat{p}_{m+1}$, the solution $w_{m+1} = w_{m+1}(\hat{p}_{m+1})$ of (3.4) exists and

$$w_{m+1} = \frac{(u_{m+1} + c)\hat{p}_{m+1}}{u_{m+1} + c + \hat{p}_{m+1}} < u_{m+1} + c.$$

Also, $w_{m+1}(\hat{p}_{m+1}) \to u_{m+1} + c$ as $\hat{p}_{m+1} \to \infty$. It is easy to see that the function $f(p) = pu/(-u+p)$ is continuous and decreasing for $p > \max\{0, u\}$. Thus, from (3.4) and the second equation of (3.5), for a sufficiently large $\hat{p}_{m+1}$ we can find $\hat{p}_m > p_m > u_m$ satisfying the first equation of (3.5) and $\hat{p}_m \to p_m$ as $\hat{p}_{m+1} \to \infty$. □

Remark 3.1. Let r_n be defined by (3.1) and assume equation (1.6) is nonoscillatory. Then Lemma 3.1 implies that we can make a perturbation to equation (1.6) at the point m by adding a new point m' between $m-1$ and m, defining or redefining $p_{m'}$, p_m, $q_{m'}$, and q_m such that $p_{m'}$ can be arbitrarily close to the original p_m, the

r_n's do not change except at $n = m'$, $r_{m'}$ can be any value we want, and the expanded equation is still nonoscillatory.

THEOREM 3.2 (another version of the telescoping principle). *Let $p_n > 0$ for $n \in \mathbb{N}$. Let (H) hold, r_n be defined by (3.1), $N = t(n)$, and the mapping T be defined as in (1.8) and (1.9). Assume equation (1.6) is nonoscillatory. Then the telescoped equation*

$$V_{N+1} = \frac{P_N V_N}{-V_N + P_N} + Q_N, \quad N = 1, 2, \ldots \tag{3.6}$$

is also nonoscillatory provided $Q_N = (Tq)_N$, $P_N = (Tp)_N$, $N \neq A_i + 1$, P_{A_i+1} is sufficiently close to p_{a_i+1}, $i = 1, \ldots, j$, and $R_N = r_n$, $N = 1, 2, \ldots$, where R_N is defined by (3.2).

Proof. According to Lemma 3.1 and Remark 3.1 we can make a perturbation to equation (1.6) at the points $a_i + 1$, $i = 1, \ldots, j$, by adding new points a_i' between a_i and $a_i + 1$ and defining or redefining p'_{a_i}, p_{a_i+1}, $q_{a_i'}$, and q_{a_i+1} such that $p_{a_i'} = P_{A_i+1}$, $r_{a_i'} = r_{b_i+1}$, the r_n''s do not change except for $n = a_i'$, $i = 1, \ldots, j$, and the expanded equation of (1.6) is still nonoscillatory.

Define

$$\bar{J} = \bigcup_{i=1}^{j} \bar{J}_i = \bigcup_{i=1}^{j} (a_i', b_i + 1]. \tag{3.7}$$

For the set $\bar{J}$, define a function $\bar{t}$ on $\mathbb{N}$ as

$$M = \bar{t}(n) = \operatorname{card}([1, n] \cap \bar{J}^c), \quad n = 1, 2, \ldots$$

and $C_i = \bar{t}(a_i')$, $i = 1, \ldots, j$. The induced transformation $\bar{T}$ on S is given as follows: for $x = \{x_n; n \in \mathbb{N}\}$ in S,

$$\bar{T}x = \bar{X} = \{\bar{X}_M;\ M \in \mathbb{N}\} \text{ with } \bar{X}_M = x_n \text{ where } \bar{t}(n) = M.$$

Consider the equation

$$W_{M+1} = \frac{\bar{P}_M W_M}{-W_M + \bar{P}_M} + \bar{Q}_M, \quad M = 1, 2, \ldots, \tag{3.8}$$

where $\bar{P} = \bar{T}p$ and $\bar{Q} = \bar{T}q$.

Since $r_{a_i'} = r_{b_i+1}$, we have $\sum_{i=a_i'+1}^{b_i+1} q_i = 0$. Noting that the expanded equation of (1.6) is nonoscillatory, by Theorem 2.3, equation (3.8) is also nonoscillatory.

Comparing (1.7) and (3.7), we see that $P_N = \bar{P}_M$, $Q_N = \bar{Q}_M$ for $N \neq A_i + 1$, $M \neq C_i + 1$, and $P_{A_i+1} = p_{a_i'} = \bar{P}_{C_i+1}$, $Q_{A_i+1} = \bar{Q}_{C_i+1}$, since $R_N = \bar{R}_M$ for $N = M = 1, 2, \ldots$. This implies that equation (3.6) is nonoscillatory and completes the proof. □

THEOREM 3.3. *Let $p_n > 0$ for $n \in \mathbb{N}$ and let (H) hold. Assume $\sum_{n=1}^{\infty} q_n$ is convergent and let $r_n^* = \sum_{i=n}^{\infty} q_i$. Let $N = t(n)$ and define the mapping T as in (1.8) and (1.9). Assume equation (1.6) is nonoscillatory. Then the telescoped equation (3.6) is also nonoscillatory provided $Q_N = (Tq)_N$, $P_N = (Tp)_N$, $N \neq A_i + 1$, P_{A_i+1} is sufficiently close to p_{a_i+1}, $i = 1, \ldots, j$, and $R_N^* = r_n^*$, $N = 1, 2, \ldots$, where $R_N^* = \sum_{i=N}^{\infty} Q_i$.*

Proof. The proof is similar to that of Theorem 3.2 and is therefore omitted. □

4. Extensions of known oscillation criteria. Here we establish some extensions of known oscillation criteria. The following assumptions are involved in the results:

$$p_n > 0 \text{ and } \sum_{n=1}^{\infty} \frac{1}{p_n} = \infty; \tag{4.1}$$

$$\sum_{n=1}^{\infty} q_n \text{ is convergent.} \tag{4.2}$$

We let

$$r_n = \sum_{i=1}^{n} q_i \tag{4.3}$$

and

$$r_n^* = \sum_{i=n}^{\infty} q_i \tag{4.4}$$

if (4.2) holds.

Result 1 (Hinton and Lewis [6]). Assume (4.1) and (4.3) hold and $\lim_{n\to\infty} r_n = \infty$. Then equation (1.1) is oscillatory.

Result 2 (Mingarelli [12]). Assume (4.1), (4.2), and (4.4) hold. If

$$\left(\sum_{n=1}^{k} \frac{1}{p_n}\right) r_k^* \geq \frac{1}{4} + \epsilon, \quad k \geq k_0 \geq 1$$

for some $\epsilon > 0$, then equation (1.1) is oscillatory.

Result 3 (Mingarelli [12]). Assume (4.1), (4.2), and (4.4) hold. If

$$\sum_{n=k}^{\infty} \frac{1}{p_n}(r_n^*)^2 \geq \left(\frac{1}{4} + \epsilon\right) r_k^*, \quad k \geq k_0 \geq 1$$

for some $\epsilon > 0$, then equation (1.1) is oscillatory.

To extend the above results, we need the definition below.

DEFINITION 4.1. *Let J be defined by* (1.7). *A sequence $p \in S$ is said to be adjusted according to J if p_{b_i+1} is replaced by p_{a_i+1}; all other $p_n's$, including p_{a_i+1}, are left unchanged for $n \neq b_i + 1$, $i = 1, 2, \ldots, j$.*

It is easy to see that $\sum_{n\notin J} 1/p_n = \infty$ for the adjusted sequence $p \in S$ is equivalent to $\sum_{n\notin J} 1/(p_{n+1}) = \infty$ for the original sequence $p \in S$. Since this concept of an adjusted sequence plays an important role below, we illustrate it for the convenience of the reader.

Illustration. Let

$$p = \{p_1, p_2, \ldots, p_{a_1}, p_{a_1+1}, \ldots, p_{b_1}, p_{b_1+1}, p_{b_1+2}, \ldots, p_{a_2}, p_{a_2+1}, \ldots, p_{b_2}, p_{b_2+1}, p_{b_2+2}, \ldots\};$$

then p adjusted according to J is given by

$$\{p_1, p_2, \ldots, p_{a_1}, p_{a_1+1}, \ldots, p_{b_1}, p_{a_1+1}, p_{b_1+2}, \ldots, p_{a_2}, p_{a_2+1}, \ldots, p_{b_2}, p_{a_2+1}, p_{b_2+2}, \ldots\}.$$

THEOREM 4.2. *Let $p_n > 0$, and let r_n be given by* (4.3) *for $n \in \mathbb{N}$. Assume that for some increasing sequence of positive integers $n_1, n_2, n_3, \ldots$,*

$$\limsup_{n\to\infty} r_n = \lim_{k\to\infty} r_{n_k} = \infty \tag{4.5}$$

and

$$\sum_{k=1}^{\infty} \frac{1}{p_{n_k+1}} = \infty. \tag{4.6}$$

Then equation (1.1) *is oscillatory.*

Proof. Assume the contrary; without loss of generality we may assume (1.6) has a solution $u \in S$ satisfying $u_n < p_n$, $n = 1, 2, \ldots$. Let $J = N \setminus \{n_1, n_2, n_3, \ldots\}$ be as in assumption (H) of (1.7). Let $N = t(n)$ and the mapping T be defined according to J. By Theorem 3.2, the telescoped equation (3.6) is also nonoscillatory and $R_k = r_{n_k}$, where $Q_N = (Tq)_N$, $P_N = (Tp)_N$ for $N \neq A_i + 1$, P_{A_i+1} is sufficiently close to P_{a_i+1}, and $R_k = \sum_{i=1}^{k} Q_i = \sum_{i=1}^{k} q_{n_i}$, $k = 1, 2, \ldots$.

Choose P_{A_i+1} so close to p_{a_i+1} that $\sum_{N=1}^{\infty} 1/P_N = \infty$. This is possible by (4.6), and for the adjusted sequence $\{p_n\}$ we can make

$$\left| \sum_{n\in J\cap[1,k]} \frac{1}{p_n} - \sum_{N=1}^{t(k)} \frac{1}{P_N} \right| \leq 1 \quad \text{for all } k.$$

Since $\lim_{N\to\infty} R_N = \lim_{k\to\infty} r_{n_k} = \infty$ by Result 1, equation (3.6) is oscillatory. This contradiction completes the proof. □

THEOREM 4.3. *Let* (4.2) *hold, r_n^* be given by* (4.4), *and I be an infinite subset of $\mathbb{N}$. Assume $p_n > 0$ for $n \in \mathbb{N}$ and*

$$\sum_{n\in I} \frac{1}{p_n} = \infty, \tag{4.7}$$

and for $k \in I$, $k \geq k_0 \geq 1$, assume that

$$\left(\sum_{n\in I\cap[1,k]} \frac{1}{p_n} \right) r_k^* \geq \frac{1}{4} + \epsilon, \tag{4.8}$$

where $\epsilon > 0$ and $\{p_n\}$ is adjusted according to $J = N \setminus I$. Then equation (1.1) *is oscillatory.*

Proof. Assume the contrary. Without loss of generality, we may assume (1.6) has a solution $u \in S$ satisfying $u_n < p_n$, $n = 1, 2, \ldots$. Let $N = t(n)$ and let the mapping T be defined according to J. By Theorem 3.2 and Remark 3.1, the telescoped equation (3.6) is also nonoscillatory, and $R_N^* = r_n^*$, where $R_N^* = \sum_{i=N}^{\infty} Q_i = \sum_{i\in I\cap[n,\infty)} q_i$ if $N = t(n)$, and P_{A_i+1} is so close to p_{a_i+1} that $\sum_{N=1}^{\infty} 1/P_N = \infty$, and

$$\left| \sum_{n\in I\cap[1,k]} \frac{1}{p_n} - \sum_{N=1}^{K} \frac{1}{P_N} \right| \leq \frac{\epsilon}{2M},$$

where $M = \max_{k\in I}\{r_k^*\} < \infty$ and $K = t(k)$. From (4.8),

$$\left(\sum_{N=1}^{K}\frac{1}{P_N}\right)R_K^* \geq \left(\sum_{n\in I\cap[1,k]}\frac{1}{p_n} - \frac{\epsilon}{2M}\right)r_k^*$$

$$\geq \left(\sum_{n\in I\cap[1,k]}\frac{1}{p_n}\right)r_k^* - \frac{\epsilon}{2} \geq \frac{1}{4} + \epsilon - \frac{\epsilon}{2} = \frac{1}{4} + \frac{\epsilon}{2}.$$

According to Result 2, equation (3.6) is oscillatory. We reach a contradiction. □

An immediate consequence of Theorem 4.3 is as follows.

COROLLARY 4.4. *Let* (4.2) *hold,* r_n *and* r_n^* *be given by* (4.3) *and* (4.4), *respectively,* $I = \{n \leq \mathbb{N};\ r_n^* \geq 0\}$, *and* $(r_n^*)_+ = \max\{r_n^*, 0\}$. *Let* $p_n > 0$ *for* $n \in \mathbb{N}$, *and assume that* (4.7) *and* (4.8)*—with* r_k^* *replaced by* $(r_k^*)_+$*—hold, where* $\epsilon > 0$ *and* $\{p_n\}$ *is adjusted according to* I. *Then equation* (1.1) *is oscillatory.*

THEOREM 4.5. *Let* (4.2) *and* (4.7) *hold where* I *is an infinite subset of* $\mathbb{N}$. *Assume* $p_n > 0$ *for* $n \in \mathbb{N}$, *and for* $k \in I$, $k \geq k_0 \geq 1$, *assume that*

$$\sum_{n\in I\cap[k,\infty)}\frac{1}{p_n}(r_n^*)^2 \geq \left(\frac{1}{4} + \epsilon\right)r_k^*,$$

where $\epsilon > 0$ *and* $p \in S$ *is adjusted according to* $J = N \setminus I$. *Then equation* (1.1) *is oscillatory.*

Proof. The proof is similar to that of Theorem 4.3 and hence is omitted. □

5. More oscillation criteria. In this section we obtain discrete analogues of oscillation criteria of Kwong and Zettl [10], [11] for second-order differential equations which are extensions of criteria obtained by Fite, Leighton, Winter, Hartman, Willett, Olech, and others. Our results are also extensions of discrete oscillation criteria obtained by Kwong, Hooker, and Patula [9], Chen and Erbe [1], [2], and Erbe and Yan [3].

The following lemma is used in the proofs.

LEMMA 5.1. *Assume that a sequence* $u \in S$ *has the property that there exists* $\alpha > 0$, $p_i > 0$, $i = 1, 2, \ldots$, *such that*

$$\text{(5.1)}\quad \text{(i)}\qquad u_{n+1} \geq \alpha + \sum_{i=1}^{n}\frac{u_i^2}{-u_i + p_i},\quad u_1 \geq \alpha, u_n < p_n \text{ for } n = 1, \ldots, k,$$

and

$$\text{(ii)}\qquad \alpha^2\sum_{n=1}^{\infty}\frac{1}{p_{i+1}(-\alpha + p_i)} > 1.$$

Then $k < \infty$.

Proof. Assume the contrary. Then $u_n < p_n$ for $n = 1, 2, \ldots$. Consider the equation

$$\text{(5.2)}\qquad v_{n+1} = \alpha + \sum_{i=1}^{n}\frac{v_i^2}{-v_i + p_i} \quad \text{with} \quad v_1 = \alpha.$$

Noting that $g(v) = v^2/(-v + p)$ is increasing for $0 \leq v < p$, by induction we have that $u_n \geq v_n \geq \alpha$ and $v_n < p_n$ for $n = 1, 2, \ldots$. From (5.2),

$$\Delta v_n = \frac{v_n^2}{-v_n + p_n}.$$

Since $v_n < p_n$ and $h(v) = v/(-v+p)$ is increasing in v for $v < p$ and $p > 0$,

$$\frac{1}{v_i} - \frac{1}{v_{i+1}} = \frac{\Delta v_i}{v_{i+1}v_i} = \frac{v_i}{v_{i+1}(-v_i+p_i)} > \frac{\alpha}{p_{i+1}(-\alpha+p_i)},$$

hence

$$\frac{1}{\alpha} - \frac{1}{v_{n+1}} = \frac{1}{v_1} - \frac{1}{v_{n+1}} = \sum_{i=1}^{n}\left(\frac{1}{v_i} - \frac{1}{v_{i+1}}\right) > \sum_{i=1}^{n}\frac{\alpha}{p_{i+1}(-\alpha+p_i)}.$$

Therefore,

$$\frac{1}{v_{n+1}} < \frac{1}{\alpha} - \alpha\sum_{i=1}^{n}\frac{1}{p_{i+1}(-\alpha+p_i)} = \frac{1}{\alpha}\left(1 - \alpha^2\sum_{i=1}^{n}\frac{1}{p_{i+1}(-\alpha+p_i)}\right), \quad n = 1,2,\ldots .$$

By condition (ii), there exists n_1 such that $1/v_{n+1} < 0$ for $n \geq n_1$, contradicting (5.2). □

The next result is for the case when r_n is large often enough.

THEOREM 5.2. *Let $p_n > 0$ for $n \in \mathbb{N}$ and r_n be defined by* (3.1). *Let $J(\lambda) = \{n \geq 2;\ r_{n-1} < \lambda\}$ and $\bar{J}(\lambda) = \{n \geq 2;\ r_{n-1} \geq \lambda\}$. Assume there exists an increasing sequence of positive numbers*

$$\lambda_1 < \lambda_2 < \cdots < \lambda_k < \cdots, \quad \lambda_k \to \infty \text{ as } k \to \infty$$

such that

(i) *there exists $n_k \in \bar{J}(\lambda_k)$ satisfying $p_{n_k} \leq \lambda_k$, $k = 1,2,\ldots$, or*

(ii) *for $k = 1,2,\ldots$, $p_n > \lambda_k$ for all $n \in \bar{J}(\lambda_k)$ and there exists $c \in (0,1)$ such that*

$$\lambda_k^2 \sum_{n\in\bar{J}(\lambda_k)} \frac{1}{p_{(n+1)^*}(-c\lambda_k + p_n)} > 1, \tag{5.3}$$

where $(n+1)^* = \min\{m \in \bar{J}(\lambda_k) :\ m \geq n+1\}$ and $p \in S$ is adjusted according to $J(\lambda_k)$.

Then equation (1.1) *is oscillatory.*

Proof. Assume the contrary and, without loss of generality, assume equation (1.6) has a solution $u \in S$ satisfying $u_n < p_n$, $n = 1,2,\ldots$. Therefore, for $n \geq 2$,

$$u_n = u_1 + r_{n-1} + \sum_{i=1}^{n-1}\frac{u_i^2}{-u_i + p_i}. \tag{5.4}$$

(i) Choose $c \in (0,1)$. Then $p_{n_k} \leq \lambda_k$ implies that $p_{n_k} \leq k\lambda_k^2 + c\lambda_k$, and hence $\lambda_k^2/(-c\lambda_k + p_{n_k}) \geq 1/k$, so

$$\lambda_k^2 \sum_{n\in\bar{J}(\lambda_k)} \frac{1}{-c\lambda_k + p_n} \geq \frac{1}{k}, \quad k = 1,2,\ldots . \tag{5.5}$$

We first show that

$$\sum_{n=1}^{\infty}\frac{u_n^2}{-u_n + p_n} = \infty. \tag{5.6}$$

By passing to a subsequence if necessary, we may assume that $\lambda_1 > -u_1/(1-c)$. From (5.4),

$$u_n > c\lambda_1 + \sum_{i=1}^{n-1} \frac{u_i^2}{-u_i + p_i} > c\lambda_1, \quad n \in \bar{J}(\lambda_1).$$

From (5.5) there exists $\ell_1 \in \mathbb{N}$ such that $\ell_1 \geq 2$ and

$$\lambda_1^2 \sum_{\bar{J}(\lambda_1)\cap[1,\ell_1]} \frac{1}{-c\lambda_1 + p_i} \geq c.$$

Then

$$\sum_{i=1}^{\ell_1} \frac{u_i^2}{-u_i + p_i} \geq \sum_{\bar{J}(\lambda_1)\cap[1,\ell_1]} \frac{u_i^2}{-u_i + p_i} \geq c^2\lambda_1^2 \sum_{\bar{J}(\lambda_1)\cap[1,\ell_1]} \frac{1}{-c\lambda_1 + p_i} \geq c^3.$$

Next we may assume that λ_2 is so large that $\lambda_2 > \max_{1\leq n\leq \ell_1 - 1} r_n$. Then $\bar{J}(\lambda_2) \subset (\ell_1, \infty)$. Repeating the above arguments we obtain $\ell_2 \in \mathbb{N}$ such that $\ell_2 \geq \ell_1 + 1$ and

$$\lambda_2^2 \sum_{\bar{J}(\lambda_2)\cap[\ell_1,\ell_2]} \frac{1}{-c\lambda_2 + p_i} \geq \frac{c}{2},$$

hence

$$\sum_{i=\ell_1+1}^{\ell_2} \frac{u_i^2}{-u_i + p_i} \geq \frac{c^3}{2}$$

as before. In general, we obtain $\ell_n \in \mathbb{N}$, $n = 1, 2, \ldots$, such that $\ell_{n+1} \geq \ell_n + 1$ and

$$\sum_{i=\ell_n+1}^{\ell_{n+1}} \frac{u_i^2}{-u_i + p_i} \geq \frac{c^3}{n+1}.$$

The divergence of $\sum_{n=1}^{\infty} u_n^2/(-u_n + p_n)$ now follows. Since $\min \bar{J}(\lambda_k) \to \infty$ as $k \to \infty$, we can choose a k large enough such that $n_1 = \min \bar{J}(\lambda_k)$ satisfies

$$u_1 + \sum_{i=1}^{n_1-1} \frac{u_i^2}{-u_i + p_i} > 0. \tag{5.7}$$

Then from (5.4), $u_n > \lambda_k$ for all $n \in \bar{J}(\lambda_k)$. In particular, $u_{n_k} > \lambda_k \geq p_{n_k}$, contradicting the assumption.

(ii) It is easy to see that (5.3) implies (5.5) if $p_n > \lambda_k$ for $n \in \bar{J}(\lambda_k)$. Then (5.6) holds. Let λ be one of the λ_n's with n large enough and $\bar{J}(\lambda) = \bigcup_{i=1}^{j} (n_i, m_i]$ such that (5.7) holds. Rewrite (5.4) as

$$u_{n+1} = r_n + \left(u_1 + \sum_{i=1}^{n_1-1} \frac{u_i^2}{-u_i + p_i}\right) + \sum_{i=n_1}^{n} \frac{u_i^2}{-u_i + p_i}.$$

Consider the telescoped equation obtained by cutting out $J(\lambda)$ from $\mathbb{N}$, letting

$$V_{N+1} = R_N + \left(u_1 + \sum_{i=1}^{n_1-1} \frac{u_i^2}{-u_i + p_i}\right) + \sum_{i=1}^{N} \frac{V_i^2}{-V_i + P_i}, \tag{5.8}$$

where $N = t(n)$ is defined by (1.8) according to $J(\lambda)$, $R_N = r_n$, $P_N = p_n$ for $N \neq N_i + 1 = t(n_i + 1)$, and P_{N_i+1} is sufficiently close to p_{n_i+1}, $i = 1, \ldots, j$, such that

$$\lambda^2 \sum_{N=1}^{\infty} \frac{1}{P_{N+1}(-c\lambda + P_N)} > 1. \tag{5.9}$$

By Theorem 3.2, equation (5.8) is nonoscillatory, i.e., $V_N < P_N$ for $N = 1, 2, \ldots$. However, from (5.8) and (5.7) we see that

$$V_{N+1} > \lambda + \sum_{i=1}^{N} \frac{V_i^2}{-V_i + P_i}.$$

This, together with (5.9), contradicts Lemma 5.1 and completes the proof. □

COROLLARY 5.3. *Let $p_n > 0$, and r_n be given by (4.3) for $n \in \mathbb{N}$. Assume that for some increasing sequence of positive integers $n_1, n_2, n_3, \ldots$, we have*

$$\limsup_{n\to\infty} r_n = \lim_{k\to\infty} r_{n_k} = \infty$$

and $p_{n_k+1} \leq r_{n_k}$, $k = 1, 2, \ldots$. Then equation (1.1) is oscillatory.

Proof. This follows from (i) of Theorem 5.2. □

COROLLARY 5.4. *Let d be any positive number less than $(1 + \sqrt{5})/2$. Assume $p_n > 0$ for $n \in \mathbb{N}$ and there exists $n \in \bar{J}(\lambda_k)$ such that $p_n \leq d\ \lambda_k$ and $p_{(n+1)^*} \leq d\ \lambda_k$, $k = 1, 2, \ldots$, where $(n+1)^*$ is defined as in Theorem 5.2 and $\{p_n\}$ is adjusted according to $J(\lambda_k)$. Then equation (1.1) is oscillatory.*

Proof. Choose $c \in (0, 1)$ such that $(c + \sqrt{c^2 + 4})/2 \geq d$. Then

$$p_{(n+1)^*}(-c\lambda_k + p_n) \leq d(-c + d)\lambda_k^2 \leq \frac{1}{4}(c + \sqrt{c^2 + 4})(-c + \sqrt{c^2 + 4})\lambda_k^2 = \lambda_k^2,$$

hence

$$\frac{\lambda_k^2}{p_{(n+1)^*}(-c\lambda_k + p_n)} \geq 1.$$

Then either (i) or (ii) of Theorem 5.2 is satisfied. □

Example 5.1. Let $p_n = 2^{n-1}$, $n = 1, 2, \ldots$, $r_n = 2^n$ if n is even, and $r_n = 0$ if n is odd. Then the condition of Corollary 5.3 is satisifed and hence equation (1.1) is oscillatory.

As far as we know, no previously known criteria apply to Example 5.1.

The next result is for the case when $\{r_n\}$ is oscillatory.

THEOREM 5.5. *Let $J(\lambda)$ and $\bar{J}(\lambda)$ be defined as in Theorem 5.2. Let $p_n > 0$ for $n \in \mathbb{N}$ and assume that*

(i) *there exists a real number λ such that*

$$\sum_{n\in\bar{J}(\lambda)} \frac{1}{p_{(n+1)^*} p_n} = \infty, \tag{5.10}$$

where $(n+1)^* = \min\{m \in \bar{J}(\lambda) : m \geq n+1\}$ and $\{p_n\}$ is adjusted according to $J(\lambda)$;

(5.11) (ii) $$\sum_{n\in J(\lambda)} \frac{(r_{n-1}-\lambda)^2}{-(r_{n-1}-\lambda)+p_n} = \infty.$$

Then equation (1.1) *is oscillatory.*

Intuitively, conditions (i) and (ii) of Theorem 5.5 on the sequence $\{r_n\}$ mean that r_n assumes values larger than or equal to λ often enough and, on the other hand, assumes values less than λ often enough.

Proof. Assume the contrary and, without loss of generality, assume (1.6) has a solution $u \in S$ satisfying $u_n < p_n$, $n = 1, 2, \ldots$. There are two possible cases depending on whether or not $u_1 + \sum_{n=1}^{\infty} u_n^2/(-u_n + p_n) - \lambda$.

In the former case, noting that (5.10) implies that $\operatorname{card}(\bar{J}(\lambda)) = \infty$, we can choose $n_1 \in \bar{J}(\lambda)$ such that

$$u_1 + \sum_{i=1}^{n_1-1} \frac{u_i^2}{-u_i + p_i} \geq -\lambda + \epsilon.$$

Rewrite (1.6) as

$$u_{n+1} = r_n + \left(u_1 + \sum_{i=1}^{n_1-1} \frac{u_i^2}{-u_i + p_i}\right) + \sum_{i=n_1}^{n} \frac{u_i^2}{-u_i + p_i}.$$

It follows that

(5.12) $$u_{n+1} \geq \epsilon + \sum_{i=n_1}^{n} \frac{u_i^2}{-u_i + p_i} \quad \text{for} \quad n+1 \in \bar{J}(\lambda),$$

and hence $u_n > \epsilon$ for $n \in \bar{J}(\lambda)$. If there exists $n^* \in \bar{J}(\lambda)$ such that $p_{n^*} \leq \epsilon$, then $u_{n^*} > p_{n^*}$, contradicting the assumption. Otherwise, as in the proof of Theorem 5.2 (ii), by using Theorem 3.2 we may assume without loss of generality that (5.12) holds for all $n \geq n_1$, and (5.10) is replaced by

$$\sum_{n=n_1}^{\infty} \frac{1}{p_{n+1}p_n} = \infty.$$

This implies that

(5.13) $$\epsilon^2 \sum_{n=n_1}^{\infty} \frac{1}{p_{n+1}(-\epsilon + p_n)} > 1.$$

Now, (5.12), (5.13), and $u_n < p_n$, $n = 1, 2, \ldots$, contradict Lemma 5.1.

In the remaining case,

$$u_1 + \sum_{n=1}^{\infty} \frac{u_n^2}{-u_n + p_n} \leq -\lambda,$$

it follows from (1.6) that $u_{n+1} \leq r_n - \lambda < 0$. In particular, for all $n \in J(\lambda)$, $u_n \leq r_{n-1} - \lambda < 0$. Since $f(u) = u^2/(-u + p)$ is decreasing for $u < 0$ and $p > 0$, we see that for $n \in J(\lambda)$,

$$\frac{u_n^2}{-u_n + p_n} \geq \frac{(r_{n-1}-\lambda)^2}{-(r_{n-1}-\lambda)+p_n}.$$

Then

$$\sum_{n=1}^{\infty} \frac{u_n^2}{-u_n + p_n} \geq \sum_{n \in J(\lambda)} \frac{u_n^2}{-u_n + p_n} \geq \sum_{n \in J(\lambda)} \frac{(r_{n-1} - \lambda)^2}{-(r_{n-1} - \lambda)^2 + p_n} = \infty.$$

This is also a contradiction. The proof is complete. □

COROLLARY 5.6. *Let $p \in S$ be bounded, $p_n > 0$ for $n \in \mathbb{N}$, r_n be given by* (3.1), *and $J(\lambda)$, $\bar{J}(\lambda)$ be defined as in Theorem* 5.2. *Assume*

(i) *there exists a real number λ such that* $\operatorname{card}(\bar{J}(\lambda)) = \infty$,

(ii) $\sum_{n \in J(\lambda)} (\lambda - r_{n-1})^2 = \infty$.

Then equation (1.1) *is oscillatory.*

Proof. It is easy to see that condition (i) implies that (5.10) holds. If $\inf\{r_n;\ n \in \mathbb{N}\} = -\infty$, then $\sup\{\lambda - r_{n-1};\ n \in J(\lambda)\} = \infty$, and hence (5.11) holds. Otherwise, $-(r_{n-1} - \lambda) + p_n \leq c$ for a positive number c. Then, by condition (ii),

$$\sum_{n \in J(\lambda)} \frac{(r_{n-1} - \lambda)^2}{-(r_{n-1} - \lambda) + p_n} \geq \frac{1}{c} \sum_{n \in J(\lambda)} (r_{n-1} - \lambda)^2 = \infty,$$

hence (5.11) holds. By Theorem 5.5, equation (1.1) is oscillatory. □

COROLLARY 5.7. *Let $p \in S$ be bounded and assume there exist two numbers $\lambda < \mu$ such that both $\bar{J}(\mu) = \{n \geq 2;\ r_{n-1} \geq \mu\}$ and $J(\lambda) = \{n \geq 2;\ r_{n-1} < \lambda\}$ have infinite cardinality. Then equation* (1.1) *is oscillatory.*

Proof. Choose $\bar{\lambda} \in (\lambda, \mu)$. Then the hypothesis of Corollary 5.6 is satisfied with $\bar{\lambda}$ in place of λ. □

COROLLARY 5.8. *Let $p \in S$ be bounded, $p_n > 0$ for $n \in \mathbb{N}$, and*

$$-\infty \leq \liminf_{n \to \infty} r_n < \limsup_{n \to \infty} r_n \leq +\infty.$$

Then equation (1.1) *is oscillatory.*

Proof. Choose λ and μ so that

$$\liminf_{n \to \infty} r_n < \lambda < \mu < \limsup_{n \to \infty} r_n.$$

Then the conclusion follows from Corollary 5.7. □

Example 5.2. Let $p_n = 1$, $n = 1, 2, \ldots$, $r_n = 0$ if n is odd, and $r_n = -1/\sqrt{n}$ if n is even. Choose $\lambda = 0$. Then $\operatorname{card}(\bar{J}(\lambda)) = \infty$, and

$$\sum_{n \in J(\lambda)} (\lambda - r_{n-1})^2 = \sum_{k=1}^{\infty} \frac{1}{2k} = \infty.$$

By Corollary 5.6, equation (1.1) is oscillatory, although in this case $\lim_{n \to \infty} r_n = 0$.

Remark 5.1. Theorems 5.2 and 5.5 can be compared with Theorems 2 and 3 of Kwong and Zettl [10] for the differential equations case.

In the difference equations case, Chen and Erbe in [1] and [2] have strong results related to those above. However, Examples 5.2 and 5.5 do not appear to be covered by the results in [1] and [2]. In contrast to these authors, we do not assume for Theorems 5.2 and 5.5 that the leading coefficient sequence p is bounded. Also, as remarked above and illustrated with Example 5.2, Theorem 5.5 covers some cases where $\lim_{n \to \infty} r_n$ exists as a finite number. On the other hand, Chen and Erbe's results cover systems which ours do not.

REFERENCES

[1] S. CHEN AND L. H. ERBE, *Riccati techniques and discrete oscillations*, J. Math. Anal. Appl., 142 (1989), pp. 468–487.

[2] ———, *Oscillation and nonoscillation for systems of self-adjoint second-order difference equations*, SIAM J. Math. Anal., 20 (1989), pp. 939–949.

[3] L. H. ERBE AND P. YAN, *Weighted averaging techniques in oscillation theory for second order difference equations*, Canad. Math. Bull., 35 (1992), pp. 61–69.

[4] L. H. ERBE AND B. G. ZHANG, *Oscillation of second order linear difference equations*, Chinese J. Math., 16 (1988), pp. 239–252.

[5] T. FORT, *Finite Difference and Difference Equation in the Real Domain*, Oxford University Press, London, 1948.

[6] D. B. HINTON AND R. T. LEWIS, *Spectral analysis of second order difference equations*, J. Math. Anal. Appl., 63 (1978), pp. 421–438.

[7] J. W. HOOKER, M. K. KWONG, AND W. T. PATULA, *Oscillatory second order linear difference equations and Riccati equations*, SIAM J. Math. Anal., 18 (1987), pp. 54–63.

[8] J. W. HOOKER AND W. T. PATULA, *Riccati type transformations for second-order linear difference equations*, J. Math. Anal. Appl., 82 (1981), pp. 451–462.

[9] M. K. KWONG, J. W. HOOKER, AND W. T. PATULA, *Riccati type transformations for second-order linear difference equations, II*, J. Math. Anal. Appl., 107 (1985), pp. 182–196.

[10] M. K. KWONG AND A. ZETTL, *Integral inequalities and second order linear oscillation*, J. Differential Equations, 45 (1982), pp. 16–33.

[11] ———, *Asymptotically constant functions and second order linear oscillation*, J. Math. Anal. Appl., 93 (1983), pp. 475–494.

[12] A. B. MINGARELLI, *Volterra–Stieltjes Integral Equations and Generalized Ordinary Differential Expressions*, Lecture Notes in Mathematics, 989, Springer-Verlag, New York, 1983.

[13] W. T. PATULA, *Growth and oscillation properties of second-order linear difference equations*, SIAM J. Math. Anal., 10 (1979), pp. 55–61.

[14] ———, *Growth, oscillation, and comparison theorems for second-order linear difference equations*, SIAM J. Math. Anal., 10 (1979), pp. 1272–1279.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 1061–1074, July 1995

013

NONTENSOR PRODUCT WAVELET PACKETS IN $L_2(\mathbb{R}^s)$*

ZUOWEI SHEN†

Abstract. Coifman and Meyer constructed univariate and tensor product bivariate orthonormal wavelet packets in [*Orthonormal wave packet bases*, preprint]. In this note, we give a general way to construct nontensor product orthonormal multivariate wavelet packets with an exponential decay. In particular, we give concrete constructions of exponentially decaying orthonormal wavelet packets and compactly supported semiorthogonal wavelet packets of two and three variables directly from the scaling function and its refinement mask.

Key words. scaling functions, wavelets, wavelet packets, box splines

AMS subject classifications. Primary, 41A58, 41A63; Secondary, 42C30

1. Introduction.

In [7], Coifman and Meyer introduced orthogonal wavelet packets for $L_2(\mathbb{R})$ and, by using tensor products, constructed orthogonal wavelet packets in $L_2(\mathbb{R}^2)$. Recently, Chui and Li [3] studied nonorthogonal wavelet packets and their dual wavelet packets in the univariate case, thus generalizing the orthogonal wavelet packets of [7]. Applications of wavelet packets in signal processing and compression can be found in [8] and [18]. Since signals, as well as images, are multidimensional, most applications are multivariate.

In this note, we will study the general theory for multivariate wavelet packets. In particular, in $L_2(\mathbb{R}^s)$ we construct nontensor product orthogonal wavelet packets with exponential decay and compactly supported semiorthogonal wavelet packets from multivariate scaling functions. Constructions of nontensor product multivariate wavelets and prewavelets from multivariate scaling functions have been studied in [1], [4], [10]–[12], and [15]–[17]. The related results in the literature can be found in these references.

Next we introduce some notation. For $\mathbb{Z}^s$, we use the abbreviation

$$\mathbb{Z}_2^s := \mathbb{Z}^s/2\mathbb{Z}^s.$$

When $\mathbb{Z}_2^s$ is used as an index set, it can be identified with the 2^s vertices of the unit cube $\{0,1\}^s$. We denote

$$\mathbb{Z}_+^s := \{z = (z_1, \ldots, z_s) \in \mathbb{Z}^s : z_i \geq 0, \quad 1 \leq i \leq s\}.$$

Recall the standard notion

$$\hat{f}(y) := \int_{\mathbb{R}^s} \exp(-iyx) f(x) dx$$

for the Fourier transform of $f \in L_2(\mathbb{R}^s)$. For $a \in \ell_2(\mathbb{Z}^s)$, its Fourier transform is denoted by

$$\hat{a}(y) := \sum_{\alpha \in \mathbb{Z}^s} a(\alpha) \exp(-iy\alpha).$$

*Received by the editors January 27, 1993; accepted for publication (in revised form) December 10, 1993. This research was supported by National Science Foundation grant DMS-9000053.

†Center for the Mathematical Sciences, University of Wisconsin-Madison, Madison, Wisconsin 53706. Present address: Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 0511.

The convolution of f and g in $L_2(\mathbb{R}^s)$ is denoted by

$$f * g(y) := \int_{\mathbb{R}^s} f(x)g(y-x)dx$$

and, for a and b in $\ell_2(\mathbb{Z}^s)$,

$$a * b(\beta) := \sum_{\alpha\in\mathbb{Z}^s} a(\beta-\alpha)b(\alpha).$$

The *symbol* of $f \in L_2(\mathbb{R}^s)$ is defined as

$$\widetilde{f}(y) := \sum_{\alpha\in\mathbb{Z}^s} |\hat{f}(y+2\pi\alpha)|^2 = \sum_{\alpha\in\mathbb{Z}^s} f * \overline{f(-\cdot)}(\alpha)\exp(-i\alpha y), \tag{1.1}$$

where the last equality is in the $L_2(\mathbb{R}^s)$ sense. If f is a compactly supported continuous function, then equality holds pointwise. More generally, the (complex) symbol of the function f is the Laurent series

$$\widetilde{f}(z) := \sum_{\alpha\in\mathbb{R}^s} f * \overline{f(-\cdot)}(\alpha)z^\alpha; \tag{1.2}$$

(1.1) equals (1.2) on the s-dimensional torus

$$\mathbb{T} := \{z : z = \exp(-iy), \quad y \in \mathbb{R}^s\}.$$

It is well known that the functions $\{f(\cdot-\alpha), \alpha \in \mathbb{Z}^s\}$ form an orthonormal sequence if and only if $\widetilde{f}(y) = 1$ a.e. $y \in \mathbb{R}^s$.

2. Orthogonal wavelet packets. Coifman and Meyer used the scaling functions with refinement masks satisfying the conditions of Theorem 3.6 in [9] to construct univariate wavelet packets. We will use multivariate scaling functions (e.g., orthogonalized box splines or box splines) and their refinement masks to construct multivariate wavelet packets in this section.

Let φ be a refinable function called a scaling function with its refinement mask a. We assume that φ has an exponential decay and its shifts form an orthonormal basis of the space

$$V := \left\{ \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\varphi(\cdot-\alpha) : c \in \ell_2(\mathbb{Z}^s) \right\}.$$

Although results in this note still hold without the restriction that φ has an exponential decay, since we are interested in constructing exponentially decaying orthonormal wavelet packets, we are content with this restriction. If φ is an exponentially decaying function, its refinement mask a decays exponentially as well.

To simplify the notation, we introduce the dilation operator $\sigma f := 2^{s/2}f(2\cdot)$, and for any closed subspace S of $L_2(\mathbb{R}^s)$, $\sigma S := \{\sigma f : f \in S\}$.

It is well known that, for an exponentially decaying scaling function φ, the corresponding sequence of spaces $\sigma^n V$, $n \in \mathbb{Z}$, form a multiresolution of $L_2(\mathbb{R}^s)$ (cf. [1], [10]).

Let W be the orthogonal complement of σV. Suppose that a_ν, $\nu \in \mathbb{Z}_2^s\backslash\{0\}$ are exponentially decaying sequences, and functions

$$\psi_\nu := \sum_{\alpha\in\mathbb{Z}^s} a_\nu(\alpha)\varphi(2\cdot-\alpha) \tag{2.1}$$

and their shifts form an orthonormal basis for W. (The functions ψ_ν, $\nu \in \mathbb{Z}_2^s \backslash \{0\}$ are called wavelets.) Then, the sequences a_ν, $\nu \in \mathbb{Z}_2^s$, with $a_0 = a$ and their Fourier transform have the following property, whose proof can be found in [10], [16].

RESULT 2.2. *The matrix*

$$U := (2^{-s/2}\hat{a}_\nu(\cdot + \pi\mu))_{\nu,\mu \in \mathbb{Z}_2^s} \quad \textit{is unitary for all} \quad y \in \mathbb{R}^s. \tag{2.3}$$

We note that the functions ψ_ν, $\nu \in \mathbb{Z}_2^s \backslash \{0\}$, defined in (2.1) by a_ν and their shifts, form an orthonormal basis for W if and only if (2.3) holds.

It is well known and easy to check that if $s = 1$, the sequences $a_0(\alpha) := a(\alpha)$ and $a_1(\alpha) := (-1)^\alpha a(\alpha + 1)$ satisfy the above conditions (cf. [9] and [13]). For $s = 2$ and 3, in the case in which φ is symmetric, Riemenschneider and Shen constructed the exponentially decaying sequences a_ν with $a_0 = a$ which satisfy condition (2.3) (cf. [15], [16]).

For the case $s > 3$, such exponentially decaying sequences a_ν, $\nu \in \mathbb{Z}_2^s$, with $a_0 = a$ and satisfying condition (2.3), can be constructed directly from the sequence a by Jia and Shen's construction of higher dimension wavelets from symmetric scaling functions (cf. [11]).

For each $\nu \in \mathbb{Z}_2^s \backslash \{0\}$, define the space $W_\nu := \{\sum_{\alpha \in \mathbb{Z}^s} c(\alpha)\psi_\nu(\cdot - \alpha), \quad c \in \ell_2(\mathbb{Z}^s)\}$; then $W = \oplus_{\nu \in \mathbb{Z}_2^s \backslash \{0\}} W_\nu$. The space $\sigma^n W$ for each $n \in \mathbb{Z}$ is called the *wavelet space*. Furthermore,

$$L_2(\mathbb{R}^s) = V \oplus \left(\bigoplus_{n \geq 0} \sigma^n W \right) = \bigoplus_{n \in \mathbb{Z}} \sigma^n W.$$

In this section, we use the scaling function φ and the sequences a_ν with $a_0 = a$ and satisfying (2.3) to construct orthogonal wavelet packets.

Define p_0 as φ, and for an arbitrary $\kappa \in \mathbb{Z}_+^s$,

$$p_\kappa(x) = p_{2\beta+\nu}(x) := \sum_{\alpha \in \mathbb{Z}^s} 2^{s/2} a_\nu(\alpha) p_\beta(2x - \alpha), \tag{2.4}$$

where $\beta \in \mathbb{Z}_+^s$ and $\nu \in \mathbb{Z}_2^s$ are the unique numbers such that $\kappa = 2\beta + \nu$. Since φ is refinable with refinement mask a, definition (2.4) is consistent with $p_0 = \varphi$. Furthermore, the functions p_ν, $\nu \in \mathbb{Z}_2^s \backslash \{0\}$, are the wavelet functions ψ_ν. Moreover, since p_0 is an exponentially decaying function and the sequences a_ν, $\nu \in \mathbb{Z}_2^s$, are exponentially decaying sequences, it is easy to prove inductively by using (2.4) that p_κ for all $\kappa \in \mathbb{Z}_+^s$ are exponentially decaying functions.

The Fourier transform of p_κ is

$$\hat{p}_\kappa(y) = \hat{p}_{2\beta+\nu}(y) = 2^{-s/2}\hat{a}_\nu(y/2)\hat{p}_\beta(y/2) \quad \text{for all } \nu \in \mathbb{Z}_2^s. \tag{2.5}$$

It can be proved inductively that for each fixed $\kappa \in \mathbb{Z}_+^s$, the functions $\{p_\kappa(y + \alpha), \quad \alpha \in \mathbb{Z}^s\}$ form an orthonormal sequence in $L_2(\mathbb{R}^s)$. This is trivially true when $\kappa = 0$, since $p_0 = \varphi$. Assuming that it is true for all $\kappa = (n_1, \ldots, n_s) \in \mathbb{Z}_+^s$ such that $|\kappa| := n_1 + \cdots + n_s < n$ for the case $|\kappa| = n$, p_κ is defined by p_β, where $\beta \in \mathbb{Z}_+^s$ is the unique one in $\mathbb{Z}_+^s$ such that $\kappa = 2\beta + \nu$ with $\nu \in \mathbb{Z}_2^s$. Since for each $|\beta| < n$, the sequence $\{p_\beta(\cdot - \alpha), \quad \alpha \in \mathbb{Z}^s\}$ forms an orthonormal sequence, i.e.,

$\sum_{\alpha\in\mathbb{Z}^s}|\hat{p}_\beta(\cdot+2\pi\alpha)|^2=1$, we have that

$$\begin{aligned}\sum_{\alpha\in\mathbb{Z}^s}|\hat{p}_\kappa(y+2\pi\alpha)|^2&=\sum_{\alpha\in\mathbb{Z}^s}2^{-s}|\hat{a}_\nu(y/2+\pi\alpha)\hat{p}_\beta(y/2+\pi\alpha)|^2\\&=2^{-s}\sum_{\mu\in\mathbb{Z}_2^s}|\hat{a}_\nu(y/2+\pi\mu)|^2\sum_{\alpha\in\mathbb{Z}^s}|\hat{p}_\beta(y/2+\pi\mu+2\pi\alpha)|^2\\&=2^{-s}\sum_{\mu\in\mathbb{Z}_2^s}|\hat{a}_\nu(y/2+\pi\mu)|^2\\&=1.\end{aligned}$$

Hence, the sequence $\{p_\kappa(\cdot-\alpha),\quad \alpha\in\mathbb{Z}^s\}$ forms a complete orthonormal basis of the closed subspace

$$P_\kappa:=\left\{f:f=\sum_{\alpha\in\mathbb{Z}^s}c(\alpha)p_\kappa(\cdot-\alpha),\quad c\in\ell_2(\mathbb{Z}^s)\right\}$$

of $L_2(\mathbb{R}^s)$. It is clear from the construction that $P_0=V$ and $P_\nu=W_\nu$ for all $\nu\in\mathbb{Z}_2^s$.

PROPOSITION 2.6. *For an arbitrary $\beta\in\mathbb{Z}_+^s$, the space σP_β can be orthogonally decomposed into spaces $P_{2\beta+\nu}$, $\nu\in\mathbb{Z}_2^s$, i.e.,*

$$\sigma P_\beta=\bigoplus_{\nu\in\mathbb{Z}_2^s}P_{2\beta+\nu}.$$

Proof. First, we prove that

$$\sigma P_\beta=\left\{f:f=\sum_{\nu\in\mathbb{Z}_2^s}\sum_{\alpha\in\mathbb{Z}^s}b_\nu(\alpha)p_{2\beta+\nu}(\cdot-\alpha),\quad b_\nu\in\ell_2(\mathbb{Z}^s)\right\}.$$

It follows from definition (2.4) that each $p_{2\beta+\nu}\in\sigma P_\beta$, $\nu\in\mathbb{Z}_2^s$. We next prove that for each $f\in\sigma P_\beta$,

$$f:=2^s\sum_{\alpha\in\mathbb{Z}^s}c(\alpha)p_\beta(2\cdot-\alpha),\quad c\in\ell_2(\mathbb{Z}^s),$$

there exist sequences $b_\nu\in\ell_2(\mathbb{Z}^s)$, such that

$$f=\sum_{\nu\in\mathbb{Z}_2^s}\sum_{\alpha\in\mathbb{Z}^s}b_\nu(\alpha)p_{2\beta+\nu}(\cdot+\alpha).\tag{2.7}$$

Taking the Fourier transforms of (2.7) and using (2.5) gives us

$$\hat{f}(y)=\hat{c}(y/2)\hat{p}_\beta(y/2)=2^{-s/2}\sum_{\nu\in\mathbb{Z}_2^s}\hat{b}_\nu(y)\hat{a}_\nu(y/2)\hat{p}_\beta(y/2).\tag{2.8}$$

Canceling $\hat{p}_\beta(y/2)$ in (2.8), we have

$$\hat{c}(y/2)=2^{-s/2}\sum_{\nu\in\mathbb{Z}_2^s}\hat{b}_\nu(y)\hat{a}_\nu(y/2).\tag{2.9}$$

For each $c \in \ell_2(\mathbb{Z}^s)$, finding such sequences b_ν, $\nu \in \mathbb{Z}_2^s$, is equivalent, for each $\hat{c}(y) \in L_2[0, 2\pi]$, to finding $\hat{b}_\nu \in L_2[0, 2\pi]$, $\nu \in \mathbb{Z}_2^s$, such that the above equation holds. Since $\hat{c}(y/2)$ and $\hat{a}_\nu(y/2)$, $\nu \in \mathbb{Z}_2^s$ are 4π-periodic, the solvability of system (2.9) is equivalent, for each $\hat{c}(y) \in L_2[0, 2\pi]$, to finding $\hat{b}_\nu \in L_2[0, 2\pi]$, which satisfies the following system of equations:

$$\hat{c}(y/2 + \pi\mu) = 2^{-s/2} \sum_{\nu \in \mathbb{Z}_2^s} \hat{b}_\nu(y)\hat{a}_\nu(y/2 + \pi\mu), \quad \mu \in \mathbb{Z}_2^s. \tag{2.10}$$

The solvability of system (2.10) in $L_2[0, 2\pi]$ for each $\hat{c}(y) \in L_2[0, 2\pi]$ follows from the facts that its coefficient matrix $U = (2^{-s/2}\hat{a}_\nu(y/2 + \pi\mu))_{\nu,\mu \in \mathbb{Z}_2^s}$ is unitary and each entry of U is measurable and bounded. Hence,

$$\sigma P_\beta = \left\{ f : f = \sum_{\nu \in \mathbb{Z}_2^s} \sum_{\alpha \in \mathbb{Z}^s} b_\nu p_{2\beta+\nu}(\cdot - \alpha), \quad b_\nu \in \ell_2(\mathbb{Z}^s), \quad \text{for all} \quad \nu \in \mathbb{Z}_2^s \right\}.$$

Finally, we show that the functions $p_\nu(\cdot - \alpha)$, $\nu \in \mathbb{Z}_2^s$, and $\alpha \in \mathbb{Z}^s$ are an orthonormal basis of σP_β by using the fact that $\tilde{p}_\kappa = 1$ for all $\kappa \in \mathbb{Z}_+^s$. For arbitrary ν_1 and ν_2 in $\mathbb{Z}_2^s$ and $j \in \mathbb{Z}^s$,

$$\begin{aligned}
\int_{\mathbb{R}^s} p_{2\beta+\nu_1}(x - j)\overline{p_{2\beta+\nu_2}(x)}dx &= \frac{1}{(2\pi)^s} \int_{\mathbb{R}^s} \exp(-ijy)\hat{p}_{2\beta+\nu_1}(y)\overline{\hat{p}_{2\beta+\nu_2}(y)}dy \\
&= \frac{1}{(2\pi)^s} \int_{\mathbb{R}^s} 2^{-s} \exp(-ijy)\hat{a}_{\nu_1}(y/2)\overline{\hat{a}_{\nu_2}(y/2)}|\hat{p}_\beta(y/2)|^2 dy \\
&= \frac{1}{(2\pi)^s} \int_{[0,4\pi]^s} 2^{-s} \exp(-ijy)\hat{a}_{\nu_1}(y/2)\overline{\hat{a}_{\nu_2}(y/2)} \sum_{\alpha \in \mathbb{Z}^s} |\hat{p}_\beta(y/2 + 2\pi\alpha)|^2 dy \\
&= \frac{1}{(2\pi)^s} \int_{[0,4\pi]^s} 2^{-s} \exp(-ijy)\hat{a}_{\nu_1}(y/2)\overline{\hat{a}_{\nu_2}(y/2)}dy \\
&= \frac{1}{(2\pi)^s} \int_{[0,2\pi]^s} 2^{-s} \exp(-ijy) \sum_{\mu \in \mathbb{Z}_2^s} \hat{a}_{\nu_1}(y/2 + \pi\mu)\overline{\hat{a}_{\nu_2}(y/2 + \pi\mu)}dy \\
&= \delta_{0,j}\delta_{\nu_1,\nu_2}.
\end{aligned}$$

Hence, the functions

$$\{p_{2\beta+\nu}(\cdot - \alpha), \quad \nu \in \mathbb{Z}_2^s, \quad \alpha \in \mathbb{Z}^s\}$$

form a complete orthonormal basis of σP_β. □

For an arbitrary $n \in \mathbb{Z}_+$, define the sets

$$nI := \{\kappa = (n_1, \ldots, n_s) \in \mathbb{Z}_+^s \backslash \{0\} : 2^{n-1} \le n_i \le 2^n - 1, \quad 1 \le i \le s\}.$$

We are ready to state the main result of this section.

THEOREM 2.11. *The functions*

$$\{p_\kappa(\cdot - \alpha), \quad \kappa \in nI, \quad \alpha \in \mathbb{Z}^s\}$$

form a complete orthonormal basis of $\sigma^n W$. *In particular, the functions* $\{p_\kappa(\cdot - \alpha), \quad \kappa \in \mathbb{Z}_+^s, \quad \alpha \in \mathbb{Z}^s\}$ *form a complete orthonormal basis of* $L_2(\mathbb{R}^s)$.

Proof. Since

$$\sigma P_0 = \bigoplus_{\nu \in \mathbb{Z}_2^s} P_\nu,$$

we have

$$\sigma P_0 \ominus P_0 = \bigoplus_{\nu \in \mathbb{Z}_2^s \backslash \{0\}} P_\nu. \tag{2.12}$$

Hence, $\sigma P_0 \ominus P_0 = W$, since $P_0 = V$ and $W = \bigoplus_{\nu \in \mathbb{Z}_2^s \backslash \{0\}} W_\nu = \bigoplus_{\nu \in \mathbb{Z}_2^s \backslash \{0\}} P_\nu$.

It can be proved inductively by using Proposition 2.6 and (2.12) that

$$\sigma^n P_0 \ominus \sigma^{n-1} P_0 = \bigoplus_{\kappa \in nI} P_\kappa. \tag{2.13}$$

Since $\sigma^n P_0 \ominus \sigma^{n-1} P_0 = \sigma^{n-1} W$, by (2.13),

$$L_2(\mathbb{R}^s) = V \oplus \left(\bigoplus_{n \geq 0} \sigma^n W \right) = P_0 \oplus \left(\bigoplus_{n>0} \left(\bigoplus_{\kappa \in nI} P_\kappa \right) \right) = \bigoplus_{\kappa \in \mathbb{Z}_+^s} P_\kappa. \tag{2.14}$$

Therefore, the fact that the functions $\{p_\kappa(\cdot - \alpha), \quad \kappa \in \mathbb{Z}_+^s, \quad \alpha \in \mathbb{Z}^s\}$ form an orthonormal basis of $L_2(\mathbb{R}^s)$ follows from the fact that the sequence $\{p_\kappa(\cdot - \alpha), \quad \alpha \in \mathbb{Z}^s\}$ forms an orthonormal basis of P_κ for each $\kappa \in \mathbb{Z}_+^s$ and (2.14). □

The functions

$$\mathcal{P} := \{2^{ns/2} p_\kappa(2^n \cdot - \alpha), \quad \kappa \in \mathbb{Z}_+^s, \quad n \in \mathbb{Z}, \quad \alpha \in \mathbb{Z}^s\}$$

are called an *orthogonal wavelet packet* of $L_2(\mathbb{R}^s)$. In the next section, we provide various ways to construct orthonormal bases of $L_2(\mathbb{R}^s)$ extracted out from $\mathcal{P}$.

3. Orthonormal bases from wavelet packets. To construct an orthogonal basis of $L_2(\mathbb{R}^s)$ from wavelets, we use the orthonormal basis of W generated by

$$\{\psi_\nu(\cdot - \alpha), \quad \nu \in \mathbb{Z}_2^s \backslash \{0\}, \quad \alpha \in \mathbb{Z}^s\}$$

and their dilations. We now have more choices of orthonormal bases for the wavelet space $\sigma^n W$ than the 2^n dilations and 2^{-n} shifts of ψ_ν, $\nu \in \mathbb{Z}_2^s \backslash \{0\}$. This gives us a chance to construct various orthonormal bases of $L_2(\mathbb{R}^s)$ according to any practical problem at hand. In this section we provide some orthonormal bases of $L_2(\mathbb{R}^s)$ by using the functions in $\mathcal{P}$. The first basis is chosen from the orthonormal basis of $\sigma^n W$ constructed in the last section.

PROPOSITION 3.1. *For each fixed $n > 0$, the functions*

$$\{2^{ks/2} p_\kappa(2^k \cdot - \alpha), \quad \kappa \in nI, \quad k \in \mathbb{Z}, \quad \alpha \in \mathbb{Z}^s\} \tag{3.2}$$

form a complete orthonormal basis of $L_2(\mathbb{R}^s)$.

Proof. Since the functions

$$\{p_\kappa(\cdot - \alpha), \quad \kappa \in nI, \quad \alpha \in \mathbb{Z}^s\} \tag{3.3}$$

form an orthonormal basis of $\sigma^n W$ by Theorem 2.11, for each k the functions

$$\{2^{ks/2} p_\kappa(2^k \cdot -\alpha), \quad \kappa \in nI, \quad \alpha \in \mathbb{Z}^s\}$$

form an orthonormal basis of $\sigma^k \sigma^n W = \sigma^{n+k} W$. For each fixed n, since

$$\bigoplus_{k\in\mathbb{Z}} \sigma^k \sigma^n W = \bigoplus_{k\in\mathbb{Z}} \sigma^{n+k} W = \bigoplus_{k\in\mathbb{Z}} \sigma^k W,$$

the functions in (3.2) are a complete orthonormal basis of $L_2(\mathbb{R}^s)$. □

It is easy to see from the proof that the above construction of a complete orthonormal basis of $L_2(\mathbb{R}^s)$ for each fixed n is done by picking an orthonormal basis given by (3.3) of the space $\sigma^n W$, instead of the known wavelet basis $\{2^{-ns/2}\psi(2^n \cdot -\alpha), \quad \alpha \in \mathbb{Z}^s\}$, then dilating the functions in (3.3) to form the functions in (3.2), which are shown to be a complete orthonormal basis of $L_2(\mathbb{R}^s)$. In the above construction, the orthonormal basis for $L_2(\mathbb{R}^s)$ varies when the integer n changes. In particular, if $n = 1$, the orthonormal basis constructed above is the one formed by the wavelets defined in (2.1). Such a generalization provides a better localization in frequency space as pointed out by Coifman and Meyer in [7]. In the construction, the integer n is fixed and the dilation k runs through $\mathbb{Z}$.

In the next construction, we allow n and k to vary simultaneously. To do this, we introduce the notion of a *disjoint covering of* $\mathbb{Z}$. A collection of pairs

$$\mathcal{J} := \{(n,k) : n \in \mathbb{Z}_+ \backslash \{0\}, \quad k \in \mathbb{Z}\}$$

is called a *disjoint covering of* $\mathbb{Z}$ if for each $\alpha \in \mathbb{Z}$, there exists a unique pair (n,k) such that $\alpha = n + k$.

THEOREM 3.4. *Let the collection of pairs $\mathcal{J}$ be a disjoint covering of $\mathbb{Z}$. Then, the functions*

$$\{2^{ks/2} p_\kappa(2^k \cdot -\alpha), \quad \kappa \in nI, \quad (n,k) \in \mathcal{J}\} \tag{3.5}$$

form a complete orthonormal basis of $L_2(\mathbb{R}^s)$.

Proof. Since the functions

$$\{p_\kappa(\cdot - \alpha), \quad \kappa \in nI, \quad \alpha \in \mathbb{Z}^s\}$$

form a complete orthonormal basis of $\sigma^n W$ by Theorem 2.11, the functions

$$\{2^{ks/2} p_\kappa(2^k \cdot -\alpha), \quad \kappa \in nI\}$$

form a complete orthonormal basis of $\sigma^{n+k} W$. Hence (3.5) forms a complete basis of $L_2(\mathbb{R}^s)$, because $\mathcal{J}$ is a disjoint covering of $\mathbb{Z}$ and $L_2(\mathbb{R}^s) = \bigoplus_{(n,k)\in\mathcal{J}} \sigma^k \sigma^n W = \bigoplus_{(n,k)\in\mathcal{J}} \sigma^{n+k} W$. □

Remark. 3.6 The results given here are still true if we use an arbitrary integer-valued dilation matrix A with the spectral radius of A^{-1} smaller than 1. Their proofs can be followed line by line with certain modifications from the proofs given here.

As an example, we construct the wavelet packets from box splines. A *box spline* can be defined on $\mathbb{R}^s$ for a given set of integer-valued matrices in $\mathbb{R}^s$ as specified by an $s \times n$ matrix Ξ via its Fourier transform

$$\widehat{M_\Xi}(y) := \prod_{\xi\in\Xi} \frac{1 - \exp(-i\xi y)}{i\xi y}. \tag{3.7}$$

If $\operatorname{rank}\Xi = s$, then M_Ξ is a nonnegative real-valued compactly supported piecewise polynomial with its support in the set

$$\Xi\square := \left\{ \sum_{\xi\in\Xi} t_\xi \xi : t_\xi \in [0,1] \right\}.$$

When $s = 1$, M_Ξ is called a *B-spline*. The relevant facts about box splines can be found in [2].

Denote

$$r(\Xi) := \min \#\{Z \subset \Xi : \quad \Xi\backslash Z \quad \text{does not span}\} - 1.$$

Then, it has been shown in box spline theory that M_Ξ is $r(\Xi) - 1$ times continuously differentiable (cf. [2]). Furthermore, M_Ξ is a symmetric function with respect to the *center*

$$c_\Xi = \sum_{\xi\in\Xi} \xi/2, \tag{3.8}$$

i.e., $M_\Xi(c_\Xi + x) = M_\Xi(c_\Xi - x)$.

Since M_Ξ is a compactly supported function, the symbol of M_Ξ, $\widetilde{M_\Xi}$, is a Laurent polynomial. If the matrix Ξ generating the box spline M_Ξ is unimodular, that is, the absolute value of the determinant of each $s \times s$ submatrix of Ξ is either 0 or 1 (cf. [2]), then its symbol never vanishes on $\mathbb{T}$. We will only use box splines generated by unimodular matrices, hence the symbol $\widetilde{M_\Xi}$ will not vanish on $\mathbb{T}$.

Define

$$\hat{\varphi} := \frac{\hat{M}_\Xi}{\widetilde{M_\Xi}^{1/2}}. \tag{3.9}$$

Since the symbol of the function φ is 1 on the torus, the function φ and its shifts form an orthonormal sequence in $L_2(\mathbb{R}^s)$.

It was proved in [15] that φ is refinable and has an exponential decay. Furthermore, its refinement mask is also an exponentially decaying sequence and can be easily calculated as follows:

(3.10)
$$a(\alpha) = \frac{2^{s/2}}{(2\pi)^s} \int_{[-\pi,\pi]^s} \left[\left(\frac{\widetilde{M_\Xi}(y)}{\widetilde{M_\Xi}(2y)} \right)^{1/2} \prod_{\xi\in\Xi} \frac{1+\exp(-i\xi y)}{2} \right] \exp(-i\alpha y)dy, \quad \alpha \in \mathbb{Z}^s.$$

Riemenschneider and Shen (cf. [15], [16]) constructed the exponentially decaying sequences a_ν from a with $a_0 = a$ as follows:

$$a_\nu(\alpha) := (-1)^{\alpha\nu} a((-1)^{2c_\Xi\nu}(\alpha + \eta(\nu))), \tag{3.11}$$

where the map $\eta : \mathbb{Z}_2^s \mapsto \mathbb{Z}_2^s$ satisfies the following conditions:

$$\eta(0) = 0, \quad (\eta(\nu) + \eta(\mu))(\nu + \mu) \quad \text{is odd}, \quad \nu \neq \mu, \quad \nu, \mu \in \mathbb{Z}_2^s. \tag{3.12}$$

Such a map was constructed in [15], [16] for $s = 1, 2, 3$. It was also remarked in [15] that such a map does not exist when $s > 3$.

It was proved in [15], by using (3.12) for sequences a_ν given by (3.11) that the matrix

$$U := (2^{-s/2}\hat{a}_\nu(\cdot + \pi\mu))_{\nu,\mu\in\mathbb{Z}_2^s}$$

is a unitary matrix. Hence, it is easy to carry out the constructions of wavelet packets from box splines and the exponentially decaying sequences a_ν given by (3.11).

Recently, Jia and Shen in [11] constructed exponentially decaying sequences a_ν for $s > 3$ with $a_0 = a$ by using Householder matrices. We note that [11] and [16] also provided a construction of exponentially decaying sequences a_ν from general symmetric scaling functions and its refinement mask.

4. Semiorthogonal wavelet packets. The orthogonal wavelet packets constructed in the previous sections are exponentially decaying but not compactly supported functions. In this section, we provide a construction of compactly supported wavelet packets. As in the construction of compactly supported prewavelets (cf. [5], [10], [14], and [16] and called semiorthogonal wavelets in [4]), we obtain the compactness of the wavelet packets by sacrificing some of the orthogonality. However, the wavelet packets constructed here still provide Riesz bases of $L_2(\mathbb{R}^s)$ and do keep some orthogonality.

Let φ be a compactly supported scaling function with a finitely supported refinement mask a. Assume that φ and its shifts form a Riesz basis of the space V.

We say that, for arbitrary function $f \in L_2(\mathbb{R}^s)$, f is *stable* if the functions $\{f(\cdot - \alpha) : \alpha \in \mathbb{Z}^s\}$ form a Riesz basis of the space

$$V_f := \left\{ \sum_{\alpha\in\mathbb{Z}^s} c(\alpha) f(\cdot - \alpha) \right\}.$$

Recall that (cf. [1], [10]) f is stable if and only if there are $0 < C_1 \le C_2 < \infty$ such that its symbol $\tilde{f}$ satisfies

$$C_1 \le \tilde{f} \le C_2. \quad \text{a.e. on } \mathbb{T}.$$

The sequence of spaces $\sigma^n V, n \in \mathbb{Z}$, generated by φ form a multiresolution of $L_2(\mathbb{R}^s)$ (cf. [1], [10]).

Let $a_\nu \in \mathbb{Z}_2^s \backslash \{0\}$ be finitely supported sequences such that the functions

$$\psi_\nu := \sum_{\alpha\in\mathbb{Z}^s} a_\nu(\alpha)\varphi(2\cdot - \alpha) \tag{4.1}$$

and their shifts form a Riesz basis for W, the orthogonal complement of σV. Then, the sequences a_ν, $\nu \in \mathbb{Z}_2^s$, with $a_0 = a$ and their Fourier transform have the following property (cf. [10], [16]).

RESULT 4.2. *The matrix* $U(y) := (\hat{a}_\nu(y + \pi\mu))_{\nu,\mu\in\mathbb{Z}_2^s}$ *has the full rank for each* $y \in \mathbb{R}^s$.

For the finitely supported sequences a_ν with $a_0 = a$ such that the functions ψ_ν given in (4.1) and their shifts form a Riesz basis for W, define p_0 as φ, and for an arbitrary $\kappa \in \mathbb{Z}_+^s$,

$$p_\kappa(x) = p_{2\beta+\nu}(x) := \sum_{\alpha\in\mathbb{Z}^s} 2^{s/2} a_\nu(\alpha) p_\beta(2x - \alpha), \tag{4.3}$$

where $\beta \in \mathbb{Z}_+^s$ and $\nu \in \mathbb{Z}_2^s$ are the unique numbers, such that $\kappa = 2\beta + \nu$. It is clear that $p_\nu = \psi_\nu$, $\nu \in \mathbb{Z}_2^s \backslash \{0\}$, where the functions ψ_ν, $\nu \in \mathbb{Z}_2^s$, are defined in (4.1). Since φ and the sequences a_ν, $\nu \in \mathbb{Z}_2^s$, are compactly supported, each function p_κ, $\kappa \in \mathbb{Z}_+^s$, is a compactly supported function and its Fourier transform can be written as

$$\hat{p}_\kappa(y) = \hat{p}_{2\beta+\nu}(y) = 2^{-s/2}\hat{a}_\nu(y/2)\hat{p}_\beta(y/2) \quad \text{for all } \nu \in \mathbb{Z}_2^s. \tag{4.4}$$

Since p_κ is a compactly supported function, we only need to check that the symbol $\tilde{p}_\kappa$ of p_κ does not vanish on the torus to prove that p_κ is stable for each $\kappa \in \mathbb{Z}_+^s$. This can be done inductively. The proof is the same as the one in §2, where we proved that p_κ given by (2.4) is an orthonormal sequence.

Since p_κ is stable, the space

$$P_\kappa := \left\{ g : g = \sum_{\alpha \in \mathbb{Z}^s} c(\alpha) p_\kappa(\cdot - \alpha), \quad c \in \ell_2(\mathbb{Z}^s) \right\}$$

is a closed subspace of $L_2(\mathbb{R}^s)$ and the functions $\{p_\kappa(\cdot - \alpha), \quad \alpha \in \mathbb{Z}^s\}$ form a Riesz basis of P_κ.

PROPOSITION 4.5. *For an arbitrary $\beta \in \mathbb{Z}_+^s$, the functions*

$$\{p_{2\beta+\nu}(\cdot - \alpha), \quad \nu \in \mathbb{Z}_2^s, \quad \alpha \in \mathbb{Z}^s\}$$

form a Riesz basis of the space σP_β.

Proof. The proof that

$$\sigma P_\beta = \left\{ f : f = \sum_{\nu \in \mathbb{Z}_2^s} \sum_{\alpha \in \mathbb{Z}^s} b_\nu(\alpha) p_{2\beta+\nu}(\cdot - \alpha), \quad b_\nu \in \ell_2(\mathbb{Z}^s) \right\}$$

is the same as the proof of the corresponding part in Proposition 2.6.

Since for each $\nu \in \mathbb{Z}_2^s$,

$$\hat{p}_{2\beta+\nu}(y) = 2^{-s/2}\hat{a}_\nu(y/2)p_\beta(y/2),$$

and the matrix $U = (2^{-s/2}\hat{a}_\nu(y/2 + \pi\mu))_{\nu,\mu \in \mathbb{Z}^s}$ has the full rank for all $y \in \mathbb{R}^s$, the sequence $\{p_{2\beta+\nu}(\cdot - \alpha), \quad \nu \in \mathbb{Z}_2^s, \quad \alpha \in \mathbb{Z}^s\}$ forms a Riesz basis of σP_β by Proposition 3.6 in [16]. □

With the above proposition, the proofs of the following two propositions are the same as the proofs of the corresponding results for orthogonal wavelet packets. In fact, one can prove these two propositions easily by following those proofs in §§2 and 3 line by line and changing "orthonormal basis" to "Riesz basis."

PROPOSITION 4.6. *The functions*

$$\{p_\kappa(\cdot - \alpha), \quad \kappa \in nI, \quad \alpha \in \mathbb{Z}^s\}$$

form a Riesz basis of $\sigma^n W$. If $n_1 \neq n_2$, then $p_{\kappa_1}(\cdot - \alpha) \perp p_{\kappa_2}$ for all $\kappa_1 \in n_1 I$, $\kappa_2 \in n_2 I$, and $\alpha \in \mathbb{Z}^s$.

The functions

$$\mathcal{P}_R := \{2^{ns/2} p_\kappa(2^n \cdot - \alpha), \quad \kappa \in \mathbb{Z}_+^s, \quad n \in \mathbb{Z}, \quad \alpha \in \mathbb{Z}^s\}$$

are called a *semiorthogonal wavelet packet* of $L_2(\mathbb{R}^s)$. We next provide various ways to construct a Riesz basis of $L_2(\mathbb{R}^s)$ extracted from $\mathcal{P}_R$.

PROPOSITION 4.7. *For each fixed $n > 0$, the functions*

$$\{2^{ks/2}p_\kappa(2^k \cdot -\alpha), \quad \kappa \in nI, \quad k \in \mathbb{Z}, \quad \alpha \in \mathbb{Z}^s\} \tag{4.8}$$

form a Riesz basis of $L_2(\mathbb{R}^s)$.

We next construct Riesz bases of $L_2(\mathbb{R}^s)$ by changing n and k simultaneously. However, this case is more complicated than the orthogonal one. As pointed out by Cohen and Daubechies in [6], semiorthogonal wavelet packets are unstable in general. To construct Riesz bases from $\mathcal{P}_R$ by changing n and k simultaneously, we need the notion of n-finite. We say that the collection of pairs

$$\mathcal{J} = \{(n,k) : n \in \mathbb{Z}_+^s \backslash \{0\}, \quad k \in \mathbb{Z}\}$$

is *n-finite*, if there is $N < \infty$ such that for all $(n,k) \in \mathcal{J}$, $n \le N$.

THEOREM 4.9. *Suppose that the collection of pairs $\mathcal{J}$ is a disjoint covering of $\mathbb{Z}$ and is n-finite. Then the functions*

$$\{2^{ks/2}p_\kappa(2^k \cdot -\alpha), \quad \kappa \in nI, \quad (n,k) \in \mathcal{J}\} \tag{4.10}$$

form a Riesz basis of $L_2(\mathbb{R}^s)$.

Proof. Since for each fixed n, the functions

$$\{p_\kappa(\cdot - \alpha), \quad \kappa \in nI, \quad \alpha \in \mathbb{Z}^s\}$$

also form a Riesz basis of $\sigma^n W$ by Proposition 4.6, the functions

$$\{2^{ks/2}p_\kappa(2^k \cdot -\alpha), \quad \kappa \in nI\}$$

form a Riesz basis of $\sigma^{n+k}W$.

Denote, for each fixed n, $\mathcal{J}_n := \{(n,k) : (n,k) \in \mathcal{J}\}$. Since $\mathcal{J}$ is n-finite, $\mathcal{J}$ is a finite disjoint union of $\mathcal{J}_n$. Furthermore, since $\mathcal{J}$ is a disjoint covering of $\mathbb{Z}$,

$$L_2(\mathbb{R}^s) = \bigoplus_n \bigoplus_{(n,k)\in\mathcal{J}_n} \sigma^{n+k}W.$$

It is clear that for each fixed n, the functions

$$\{2^{ks/2}p_\kappa(2^k \cdot -\alpha), \quad \kappa \in nI \quad (n,k) \in \mathcal{J}_n\}$$

form a Riesz basis of $\bigoplus_{(n,k)\in\mathcal{J}_n} \sigma^{n+k}W$. Therefore, since $\mathcal{J}$ is n-finite, the functions in (4.10) form a Riesz basis of $L_2(\mathbb{R}^s)$. □

If $\mathcal{J}$ is a disjoint covering of $\mathbb{Z}$, the functions $\{(2^{-s}2^{1-n}/\|\tilde{p}_\kappa\|_\infty^{1/2})p_\kappa(2^k \cdot -\alpha) : \kappa \in nI, (n,k) \in \mathcal{J}, \alpha \in \mathbb{Z}^s\}$ form a Bessel basis for $L_2(\mathbb{R}^s)$. Recall that a set of functions $f_\alpha \in L_2(\mathbb{R}^s)$, $\alpha \in X$, where X is a countable set, is called a *Bessel basis* of $L_2(\mathbb{R}^s)$ if $L_2(\mathbb{R}^s)$ is the closure of the space

$$\left\{ g : g = \sum_{\alpha\in X} b(\alpha) f_\alpha, \quad b \in \ell_0(X) \right\},$$

where ℓ_0 is the space of the finitely supported sequences on X and, furthermore, there is a constant C such that, for an arbitrary sequence $c \in \ell_2(X)$,

$$\left\| \sum_{\alpha \in X} c(\alpha) f_\alpha \right\|_2 \le C \|c\|_2 .$$

PROPOSITION 4.11. *Suppose that $L_2(\mathbb{R}^s) = \oplus_{n\in\mathbb{Z}} W_n$, where W_n is the closed subspace of $L_2(\mathbb{R}^s)$. If the functions $f_{n,k}$, $k \in \mathbb{Z}$, form a Bessel basis of W_n and there is a constant C such that for an arbitrary n and an arbitrary $c \in \ell_2(\mathbb{Z})$,*

$$\left\| \sum_{k \in \mathbb{Z}} c(\alpha) f_{n,k} \right\|_2 \le C \|c\|_2 , \tag{4.12}$$

then the functions

$$\{f_{n,k}, \quad n \in \mathbb{Z}, \quad k \in \mathbb{Z}\} \tag{4.13}$$

form a Bessel basis of $L_2(\mathbb{R}^s)$.

Proof. Let

$$H := \left\{ a = (a_n)_{n\in\mathbb{Z}} : a_n \in \ell_2(\mathbb{Z}), \quad \text{with} \quad \sum_{n\in\mathbb{Z}} \|a_n\|_2^2 < \infty \right\}.$$

For arbitrary $a = (a_n)_{n\in\mathbb{Z}} \in H$ and $b = (b_n)_{n\in\mathbb{Z}} \in H$, define

$$\langle a, b \rangle := \sum_{n\in\mathbb{Z}^s} \langle a_n, b_n \rangle;$$

then H is a Hilbert space isometric to $\ell_2(X)$, where $X = \cup\mathbb{Z}$.

For an arbitrary $a = (a_n)_{n\in\mathbb{Z}} \in H$ and the function $g := \sum_{n\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} a_n(k) f_{n,k}$,

$$\begin{aligned} \|g\|_2^2 &= \sum_{n\in\mathbb{Z}} \left\| \sum_{k\in\mathbb{Z}} a_n(k) f_{n,k} \right\|_2^2 \\ &\le \sum_{n\in\mathbb{Z}} C^2 \|a_n\|_2^2 \\ &= C^2 \|a\|_2^2 . \end{aligned} \tag{4.14}$$

Hence, (4.13) is a Bessel basis of $L_2(\mathbb{R}^s)$ by (4.14) and $L_2(\mathbb{R}^s) = \bigoplus_{n\in\mathbb{Z}} W_n$. □

THEOREM 4.15. *Let the collection of pairs $\mathcal{J}$ be a disjoint covering of $\mathbb{Z}$; then the functions*

$$\left\{ 2^{-s} 2^{1-n} \frac{1}{\|\tilde{p}_\kappa\|_\infty^{1/2}} 2^{ks/2} p_\kappa(2^k \cdot - \alpha), \quad \kappa \in nI, \quad (n.k) \in \mathcal{J} \right\} \tag{4.16}$$

form a Bessel basis of $L_2(\mathbb{R}^s)$.

Proof. Since for each fixed n the functions

$$\{p_\kappa(\cdot - \alpha), \quad \kappa \in nI, \quad \alpha \in \mathbb{Z}^s\}$$

form a Riesz basis of $\sigma^n W$ by Proposition 4.6, the functions

$$\{2^{ks/2} p_\kappa(2^k \cdot - \alpha), \quad \kappa \in nI\}$$

form a Riesz basis of $\sigma^{n+k} W$.

Since $\mathcal{J}$ disjointly covers $\mathbb{Z}$, $L_2(\mathbb{R}^s) = \bigoplus_{(n,k)\in\mathcal{J}} \sigma^{n+k} W$. To prove the functions in (4.16) form a Bessel basis of $L_2(\mathbb{R}^s)$, we note that the symbol of

$$\left\{ \frac{1}{\|\tilde{p}_\kappa\|_\infty^{1/2}} p_\kappa, \quad \kappa \in \mathbb{Z}_+ \right\}$$

is at most one. Therefore, the functions in (4.16) form a Bessel basis of $L_2(\mathbb{R}^s)$ by Proposition 4.11. □

Let M_Ξ be the box spline defined in (3.8) with its symbol not vanishing on $\mathbb{T}$. Then, function M_Ξ is refinable with the finitely supported refinement mask a, whose Fourier transform is

$$\hat{a}(y) = 2^{s/2} \prod_{\xi\in\Xi} \frac{1 + \exp(-i\xi y)}{2}. \tag{4.17}$$

On the torus, the symbol of M_Ξ can be written as

$$\widetilde{M_\Xi} = \sum_{\alpha\in\mathbb{Z}^s} (M_\Xi * M_\Xi(-\cdot))(\alpha) \exp(-i\alpha\cdot).$$

Since the box spline M_Ξ is compactly supported, $\{(M_\Xi * M_\Xi(-\cdot))(\alpha)\}_{\alpha\in\mathbb{Z}^s}$ is a finitely supported sequence. Therefore, the sequence

$$b := d * a, \tag{4.18}$$

where $d(\alpha) := (M_\Xi * M_\Xi(-\cdot))(\alpha)$, is finitely supported.

The sequences a_ν, $\nu \in \mathbb{Z}_2^s$, are defined by the sequences a, b given in (4.17), (4.18), and the map η satisfying (3.12) as follows:

$$a_0 := a,$$

and for each $\nu \in \mathbb{Z}_2^s \backslash \{0\}$,

$$a_\nu(\alpha) := (-1)^{\alpha\nu} b((-1)^{2c_\Xi \nu}(\alpha + \eta(\nu))). \tag{4.19}$$

Since the sequences a and b are finitely supported, the sequences a_ν, $\nu \in \mathbb{Z}_2^s$, are finitely supported. Furthermore, the functions $\{\psi_\nu(\cdot - \alpha), \quad \nu \in \mathbb{Z}_2^s \backslash \{0\}, \quad \alpha \in \mathbb{Z}^s\}$ defined as (4.1) with $\varphi = M_\Xi$ and a_ν given by (4.19) are compactly supported and form a Riesz basis of $\sigma V \ominus V = W$ (cf. [16]). Therefore, we can construct semiorthogonal compactly supported wavelet packets from box splines as given by (4.3).

Acknowledgments. The author thanks Carl de Boor, Ingrid Daubechies, Kirk Haller, and Sherman Riemenschneider for their comments. He also wishes to thank the referee for his suggestion to state the results here in this general setup.

REFERENCES

[1] C. DE BOOR, R. DEVORE, AND A. RON, *On the construction of multivariate (pre)wavelets*, Constr. Approx. (special issue), 9 (1993), pp. 123–166.

[2] C. DE BOOR, K. HÖLLIG, AND S. D. RIEMENSCHNEIDER, *Box splines*, Springer-Verlag, New York, 1993.

[3] C. K. CHUI AND C. LI, *Nonorthogonal wavelet packets*, SIAM J. Math. Anal., 24 (1993), pp. 712–738.

[4] C. K. CHUI, J. STÖCKLER, AND J. D. WARD, *Compactly supported box spline wavelets*, Approx. Theory Appl., 8 (1992), pp. 77–100.

[5] C. K. CHUI AND J. Z. WANG, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–915.

[6] A. COHEN AND I. DAUBECHIES, *On the instability of arbitrary biorthogonal wavelet packets*, SIAM J. Math. Anal., 24 (1993), pp. 1340–1354.

[7] R. R. COIFMAN AND Y. MEYER, *Orthonormal wave packet bases*, preprint.

[8] R. R. COIFMAN, Y. MEYER, S. QUAKE, AND M. V. WICKERHAUSER, *Signal processing and compression with wave packets*, in Progress in Wavelet Analysis and Application, Y. Meyer and S. Roques, eds., Editions Frontieres, Gif-sur-Yvette, France, 1993, pp. 77–93.

[9] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[10] R. Q. JIA AND C. A. MICCHELLI, *Using the refinement equation for the construction of pre-wavelets II: Powers of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.

[11] R. Q. JIA AND Z. SHEN, *Multiresolution and wavelets*, Proc. Edinburgh Math. Soc. (2), 37 (1994), pp. 271–300.

[12] R. A. LORENTZ AND W. R. MADYCH, *Wavelets and generalized box splines*, Appl. Anal., 44 (1992), pp. 51–76.

[13] S. G. MALLAT, *Multiresolution approximations and wavelet orthonormal bases of* $L^2(\mathbb{R})$, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[14] C. A. MICCHELLI, *Using the refinement equation for the construction of pre-wavelets*, Numer. Algorithms, 1 (1991), pp. 75–116.

[15] S. D. RIEMENSCHNEIDER AND Z. SHEN, *Box splines, cardinal series and wavelets*, in Approximation Theory and Functional Analysis, C. K. Chui, ed., Academic Press, New York, 1991, pp. 133–149.

[16] ———, *Wavelets and pre-wavelets in low dimensions*, J. Approx. Theory, 71 (1992), pp. 18–38.

[17] J. STÖCKLER, *Multivariate wavelets*, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, New York, 1992, pp. 325–355.

[18] M. V. WICKERHAUSER, *Acoustic signal compression with wavelet packets*, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, New York, 1992, pp. 679–700.

SIAM J. MATH. ANAL.
Vol. 26, No. 4, pp. 1075–1087, July 1995

014

ASYMPTOTIC REGULARITY OF COMPACTLY SUPPORTED WAVELETS*

HANS VOLKMER†

Abstract. We study the asymptotic regularity of orthonormal bases of compactly supported wavelets when their support width tends to infinity. We construct sequences of wavelets whose ratio of regularity to support width is greater than that in previously known examples.

Key words. two-scale difference equations, wavelets, Sobolev regularity

AMS subject classifications. 26A16, 42C15, 94A11

1. Introduction.

There is a well-known method for constructing compactly supported wavelet bases in $L^2(\mathbf{R})$. We start with the two-scale difference equation

$$\phi(x) = \sum_{k=M}^{N} 2c_k \phi(2x-k), \tag{1.1}$$

where, at this point, the only condition on the complex numbers c_k is that $\sum c_k = 1$. In order to solve this equation we first define the trigonometric polynomial

$$m_0(z) = \sum_{k=M}^{N} c_k e^{ikz}. \tag{1.2}$$

Then the inverse Fourier transform ϕ of the entire function

$$A(z) = \prod_{j=1}^{\infty} m_0(2^{-j}z) \tag{1.3}$$

is a solution of (1.1). In general, ϕ is a distribution with compact support contained in the interval $[M,N]$. If ϕ (or equivalently A) is in $L^2(\mathbf{R})$ and m_0 satisfies Cohen's criterion [12, Def. 5.2], then ϕ is a scaling function of multiresolution analysis. Cohen's criterion is satisfied if m_0 does not have zeros in $[-\pi/2, \pi/2]$. Then a standard definition leads to the associated wavelet and the corresponding wavelet basis of $L^2(\mathbf{R})$.

In this paper we study the (Sobolev) *regularity index* of scaling functions and wavelets obtained in the way described above. By the regularity index of a function defined on $\mathbf{R}$ we mean the supremum of all $s \in \mathbf{R}$ such that the considered function belongs to the Sobolev space H^s. The regularity index of wavelets was studied in [1]–[8], [11]–[13]. Since wavelets have the same regularity index as the scaling function from which they are derived [12, Prop. 10.3], it is sufficient to investigate the regularity index of the scaling function ϕ or, equivalently, the growth of the entire function A for $z \to \pm\infty$. Our first problem is thus quite simple to formulate.

PROBLEM 1. *Given a trigonometric polynomial m_0 with $m_0(0) = 1$, find the behavior of the entire function A for $z \to \pm\infty$.*

* Received by the editors May 14, 1993; accepted for publication (in revised form) December 10, 1993.

† Department of Mathematical Sciences, University of Wisconsin–Milwaukee, P.O. Box 413, Milwaukee, Wisconsin 53201.

In §§2 and 3 we give a method for handling this problem.

We are interested in finding scaling functions as regular as possible under given constraints. For example, it is known that the most regular scaling function with a given support width S (i.e., the length of the interval $[M, N]$) has regularity index $S-1$. It is given by a basic spline and m_0 is a power of $(1+e^{iz})/2$ [12, p. 1532]. However, the following problem is open (see [12, p. 1542]).

PROBLEM 2. *Find the most regular "orthonormal" scaling functions with a given support width.*

By an orthonormal scaling function we mean a scaling function ϕ such that its integer translates $\phi(x-n)$ form an orthonormal system in $\mathrm{L}^2(\mathbf{R})$. The corresponding condition for m_0 is

$$|m_0(z)|^2 + |m_0(z+\pi)|^2 = 1, \qquad z \in \mathbf{R}. \tag{1.4}$$

In this paper, we attack this problem when the support width of the scaling function tends to infinity. We know that the regularity index can grow only linearly with the support width. It is therefore useful to consider the ratio of the regularity index to the support width. Let us call this quotient the *regularity ratio*. We then arrive at the following problem.

PROBLEM 3. *Find sequences ϕ_n of orthonormal scaling functions whose support width tends to infinity as $n \to \infty$ such that the limit (if it exists, otherwise use the* lim inf*) of their regularity ratios is as large as possible.*

Let us call this limit (or lim inf) the *asymptotic regularity ratio* of the sequence of scaling functions. We should mention that Problems 1 and 2 depend on the type of regularity (Sobolev, Hölder, etc.) that we use, but Problem 3 is independent of the notion of regularity.

In the well-known example of Daubechies's wavelets, the asymptotic regularity ratio equals $0.5\log_4(4/3) = 0.1037\ldots$; see [13]. We prove this result again in §5. As the main result of this paper, we construct sequences of orthonormal scaling functions with an asymptotic regularity ratio greater than $0.5\log_4(4/3)$ in §6.

2. Regularity bounds. Let m_0 be a trigonometric polynomial with $m_0(0) = 1$, and let A be the corresponding entire function (1.3). We define 2π-periodic functions

$$f_m(z) = \left(\prod_{j=0}^{m-1} |m_0(2^j z)| \right)^{1/m}, \qquad m \in \mathbf{N}. \tag{2.1}$$

We note the equation

$$|A(2^m z)| = |A(z)| f_m(z)^m \tag{2.2}$$

It is known [1] that the sequence λ_m of the maximum norms of the functions f_m,

$$\lambda_m = \max\{f_m(z) : z \in [-\pi, \pi]\}, \tag{2.3}$$

converges to its infimum. This follows from the submultiplicativity of the sequence λ_m^m. Another proof is given in the next section (Theorem 3.2). We denote the limit of the sequence λ_m by $\lambda = \lambda(m_0)$. We have $\lambda \geq 1$ because $m_0(0) = 1$.

THEOREM 2.1. *The function $|z|^s |A(z)|$ is bounded for $z \to \pm\infty$ if $s < -\log_2 \lambda$.*

Proof. Let $s < -\log_2 \lambda \leq 0$. Then there exists $k_0 \in \mathbf{N}$ such that $\lambda_k \leq 2^{-s}$ for $k \geq k_0$. Now (2.2) shows that

$$|A(z)| \leq |A(2^{-k} z)| 2^{-sk} \text{ for all } z \in \mathbf{R}, k \geq k_0.$$

If we choose $2^k \le |z| < 2^{k+1}$, then we obtain

$$|z|^s |A(z)| \le C \text{ for all } |z| \ge 2^{k_0},$$

where C is the maximum of $|A(w)|$ for $|w| \le 2$. □

Usually, the above result is applied only if m_0 does not have zeros at π. In the general case we proceed as follows. Let π be a zero of order L of m_0. Then there is a trigonometric polynomial $\tilde{m}_0$ such that

$$m_0(z) = \left(\frac{1+e^{iz}}{2}\right)^L \tilde{m}_0(z) \tag{2.4}$$

and

$$A(z) = \left(\frac{e^{iz}-1}{iz}\right)^L \tilde{A}(z), \tag{2.5}$$

where $\tilde{A}$ is defined as in (1.3) with $\tilde{m}_0$ in place of m_0. Together with Theorem 2.1 this equation implies the following result.

THEOREM 2.2. *If m_0 has the form (2.4), then the function $|z|^s|A(z)|$ is bounded for $z \to \pm\infty$ if $s < L - \log_2 \lambda(\tilde{m}_0)$. In particular, the regularity index of the inverse Fourier transform ϕ of A is at least*

$$L - \log_2 \lambda(\tilde{m}_0) - \frac{1}{2}.$$

The final statement of the above theorem also follows from [12, Prop. 9.5 and 9.7]. The same reference also implies the following partial converse of the above theorem.

THEOREM 2.3. *If the trigonometric polynomial $\tilde{m}_0$ does not vanish at $z = \pi$ and satisfies Cohen's criterion, then the regularity index of the scaling function ϕ associated with m_0 is at most*

$$L - \log_2 \lambda(\tilde{m}_0).$$

We remark that the regularity index of ϕ can be calculated more precisely by other methods; see [6], [7] for Sobolev regularity and [11] for Hölder regularity. However, these methods tend to become tedious if the support width is large. The above estimates are appropriate for the study of the asymptotic behavior of the regularity index when the support width tends to infinity.

3. Geometric means. In the previous section we saw an example of the following more general problem: let

$$T : [a,b] \to [a,b] \tag{3.1}$$

be a given transformation of the interval $[a,b]$ into itself. Then the iterates T^m of T are well defined. As usual, T^0 denotes the identity transformation. Let

$$f : [a,b] \to [0,\infty)$$

be a given bounded nonnegative function on $[a,b]$. We define the geometric means

$$f_m(x) = \left(\prod_{j=0}^{m-1} f(T^j x)\right)^{1/m}, \qquad x \in [a,b]$$

and their suprema

$$\lambda_m = \sup\{f_m(x) : x \in [a,b]\}. \tag{3.2}$$

Then the problem is to find the behavior of the sequence λ_m as m tends to infinity.

In the situation in §2 we take $[a,b] = [-\pi,\pi]$ and $T = U$, where

$$U(x) = \begin{cases} 2x & \text{if} \quad -\pi/2 \le x \le \pi/2, \\ 2x - 2\pi & \text{if} \quad \pi/2 < x \le \pi, \\ 2x + 2\pi & \text{if} \quad -\pi \le x < -\pi/2, \end{cases} \tag{3.3}$$

and $f(x) = |m_0(x)|$ or $f(x) = |\tilde{m}_0(x)|$.

For another closely related example, we take $[a,b] = [0,1]$ and $T = W$, where

$$W(t) = 4t(1-t). \tag{3.4}$$

If f is an even function on $[-\pi,\pi]$, then the numbers λ_m formed with respect to the transformation $T = U$ coincide with those formed with respect to the function $g(t) = f(\sin^2(t/2))$ and the transformation $T = W$.

The above problem is related to ergodic theory. In ergodic theory one usually considers arithmetic means but the geometric means can be written as arithmetic means using the logarithm. In ergodic theory one assumes that T is measure-preserving and ergodic, which is the case for the transformation $T = U$ with respect to the Lebesgue measure. Then the individual ergodic theorem [9, p. 18] states that the sequence of functions f_m converges almost everywhere to a constant function if $\ln f$ is Lebesgue-integrable. However, this and similar results do not help much in solving our problem because we consider all $x \in [a,b]$ in the definition (3.2).

We further remark that λ_m^m is the operator norm of the operator S^m, where $S : B[a,b] \to B[a,b]$ is the operator defined by $(Sg)(x) = f(x)g(Tx)$ and $B[a,b]$ denotes the Banach space of bounded functions on $[a,b]$ equipped with the supnorm. By Gelfand's formula for the spectral radius of a bounded linear operator, we see that the sequence λ_m converges to the spectral radius of the operator S. However, we will not use this fact in what follows.

We now return to the general situation described above. In order to investigate the sequence λ_m, it is useful to consider the functions

$$g_m(x) = \min\{f_k(x) : k = 1,\dots,m\}, \qquad x \in [a,b], \tag{3.5}$$

and their suprema

$$\mu_m = \sup\{g_m(x) : x \in [a,b]\}. \tag{3.6}$$

The sequence μ_m is nonincreasing and $\mu_m \le \lambda_m$. The numbers λ_m and μ_m are also related in the following way.

LEMMA 3.1. *Assume that $\mu_k > 0$ for a fixed $k \in \mathbf{N}$. Then*

$$\lambda_m \le c^{1/m}\mu_k \text{ for all } m \in \mathbf{N},$$

where

$$c = \max\left\{\left(\frac{\lambda_j}{\mu_k}\right)^j : j = 1,\dots,k\right\}.$$

is a constant independent of m.

Proof. We prove the above inequality by induction on m. The inequality is true for all $m = 1, \dots, k$ by the definition of c. We now assume that the inequality holds for $m-k+1, \dots, m$ in place of m, and we are going to show that the inequality also holds with $m+1$ in place of m. Let $x \in [a, b]$ be arbitrary. By the definition of g_k, there is $l \in \{1, \dots, k\}$ such that $g_k(x) = f_l(x)$. The functional equation

$$f_{m+1}(x)^{m+1} = f_l(x)^l f_{m+1-l}(T^l x)^{m+1-l}$$

shows that

$$f_{m+1}(x)^{m+1} \le \mu_k^l (\lambda_{m+1-l})^{m+1-l}.$$

Now the induction hypothesis implies

$$f_{m+1}(x)^{m+1} \le \mu_k^l c \mu_k^{m+1-l} = c\mu_k^{m+1}.$$

Since x is arbitrary, this proves the desired result. □

If $\mu_k = 0$, then the lemma remains true if we replace μ_k by $\epsilon > 0$.

We now obtain the following result.

THEOREM 3.2. *The sequence* λ_m *is convergent. Its limit* λ *satisfies*

$$\lambda = \inf_m \lambda_m = \inf_m \mu_m.$$

Proof. The sequence μ_m converges to its infimum because it is nonincreasing and nonnegative. Lemma 3.1 and the subsequent remark show that

$$\limsup_{m \to \infty} \lambda_m \le \mu_k \text{ for all } k \in \mathbf{N}.$$

Since $\mu_m \le \lambda_m$ for all m, this proves that the sequence λ_m converges to the same limit as the sequence μ_m. It is then clear that the limit λ equals the infimum of both of the sequences λ_m and μ_m. □

The above-mentioned problem can now be reformulated as follows: given the transformation T and the function f, find the limit λ of the sequence λ_m. Usually, the calculation of λ is difficult but we can always estimate λ. By Theorem 3.2, each λ_m, μ_m is an upper bound for λ. The upper bounds μ_m usually lead to much better estimates than the upper bounds λ_m. For example, in the case of the Daubechies wavelets, the sequence μ_m becomes constant beginning with μ_2, so that $\lambda = \mu_2$, whereas the bound λ_m never gives the exact value of λ (see §5). We remark that the idea to use the numbers μ_m to estimate λ also appears in a somewhat different form in [5, Lem. 7.1.6].

The following theorem provides lower bounds for λ.

THEOREM 3.3. *Let* x *be a fixed point of* T^k, *where* $k \in \mathbf{N}$. *Then*

$$\lambda \ge f_k(x).$$

Proof. By assumption, $f_{mk}(x) = f_k(x)$ for all $m \in \mathbf{N}$. This implies $\lambda_{mk} \ge f_k(x)$ for all $m \in \mathbf{N}$. As $m \to \infty$, we obtain $\lambda \ge f_k(x)$. □

In the special case $T = W$, we see that $3/4$ is a fixed point of T which yields

$$\lambda \ge f(3/4) \text{ if } T = W. \tag{3.7}$$

4. Constructing scaling functions. Let p be a polynomial of degree n which satisfies

$$p(x) + p(1-x) = 1 \text{ and } p(0) = 1. \tag{4.1}$$

Furthermore, let $p(x)$ be nonnegative for $0 \le x \le 1$. Since $p(\sin^2(z/2))$ is an even nonnegative trigonometric polynomial, there exists a trigonometric polynomial

$$m_0(z) = \sum_{k=0}^{n} c_k e^{ikz}$$

with real coefficients such that

$$|m_0(z)|^2 = p(\sin^2(z/2)) \text{ and } m_0(0) = 1. \tag{4.2}$$

This follows from a well-known theorem of F. Riesz [10, p. 81]. The trigonometric polynomial m_0 satisfies the functional equation (1.4) because p satisfies (4.1). We note that m_0 is not uniquely determined by (4.2), but this does not play a role here because we are only concerned with $|m_0(z)|$. If, in addition, m_0 satisfies Cohen's criterion, then the inverse Fourier transform ϕ of the infinite product (1.3) solves the two-scale difference equation (1.1), and ϕ is a scaling function of multiresolution analysis leading to an orthonormal wavelet basis. The support width of this scaling function is equal to the degree of the polynomial p. The regularity index of ϕ can be estimated in the following way.

THEOREM 4.1. *Let p be a polynomial solution of* (4.1) *that is nonnegative on* $[0,1]$. *Let p have a zero of exact order L at $x = 1$, and let q be the polynomial defined by*

$$p(x) = (1-x)^L q(x).$$

Let λ be the limit of the sequence λ_m defined in §3 with respect to $f = q$ and the transformation $W(x) = 4x(1-x)$.

Then the regularity index σ of an orthonormal scaling function ϕ associated with the polynomial p as introduced above satisfies

$$L - \log_4 \lambda - \frac{1}{2} \le \sigma \le L - \log_4 \lambda.$$

Proof. Let m_0 be a trigonometric polynomial satisfying (4.2). Then there is a trigonometric polynomial $\tilde{m}_0$ such that

$$m_0(z) = \left(\frac{1+e^{iz}}{2}\right)^L \tilde{m}_0(z) \text{ and } |\tilde{m}_0(z)|^2 = q(\sin^2(z/2)).$$

Now λ is the square of the number $\lambda(\tilde{m}_0)$ used in §2. Therefore, the statement of the theorem follows from Theorems 2.2 and 2.3. □

We solve equation (4.1) in the following way.

PROPOSITION 4.2. *The polynomial solutions of* (4.1) *are given by*

$$p(x) = \frac{\int_x^1 s(W(t))dt}{\int_0^1 s(W(t))dt}, \tag{4.3}$$

where s is any polynomial such that the denominator of the fraction is nonzero.

Proof. A polynomial p satisfies (4.1) if and only if $p'(x) - p'(1-x) = 0, p(0) = 1$, and $p(1) = 0$. The equation for p' means that $p'(x)$ is a polynomial in $(x - 1/2)^2$ or, equivalently, in $W(x)$. This shows that the polynomials given by (4.3) solve (4.1) and, conversely, each solution is of this form. □

We now construct sequences of orthonormal scaling functions in the following way. We start with an arbitrary real polynomial r (not identically zero) and define, for each $n \in \mathbf{N}$,

$$p_n(x) = \frac{1}{\alpha_n} \int_x^1 r(W(t))^n dt, \text{ where } \alpha_n = \int_0^1 r(W(t))^n dt \tag{4.4}$$

if $\alpha_n \neq 0$. Thus p_n is the polynomial (4.3) when $s(y) = r(y)^n$. By Proposition 4.2, this polynomial satisfies (4.1). If n is even then $\alpha_n > 0$, p_n is nonnegative on $[0, 1]$, and Cohen's criterion is satisfied because $p_n(x) > 0$ for $0 \leq x < 1$. If r itself is nonnegative on $[0, 1]$, then, of course, this is true for all n. For each even n, let ϕ_n be the scaling function associated with the trigonometric polynomial m_0 derived from $p_n(\sin^2(z/2))$ using (4.2).

Our goal is to investigate the asymptotic regularity ratio of these sequences of scaling functions. We do this in §6, but let us first consider a well-known special case, the sequence of scaling functions introduced by Daubechies [3].

5. Daubechies's scaling functions. For each $n \in \mathbf{N}$, we consider the polynomial

$$p_n(x) = \frac{1}{\alpha_n} \int_x^1 W(t)^n dt, \tag{5.1}$$

where

$$\alpha_n = \int_0^1 W(t)^n dt = 4^n \frac{(n!)^2}{(2n+1)!}. \tag{5.2}$$

This is the special case of definition (4.4) when $r(y) = y$. The polynomial p_n is of degree $2n + 1$ and has a zero at $x = 1$ of order $n + 1$. Let q_n be the polynomial of degree n defined by

$$p_n(x) = (1 - x)^{n+1} q_n(x). \tag{5.3}$$

If we differentiate this equation and compare it with $p_n' = -W(x)^n/\alpha_n$, then we obtain

$$(1 - x)q_n'(x) = (n + 1)q_n(x) - (2n + 1)\binom{2n}{n} x^n. \tag{5.4}$$

We use this differential equation to show that

$$q_n(x) = \sum_{j=0}^{n} \binom{n+j}{j} x^j.$$

In fact, we easily verify that the right-hand side satisfies the same linear first-order differential (5.4) as q_n and the two solutions agree for $x = 0$.

The polynomials q_n are the same as those considered in [3, (4.13)]. Daubechies's scaling functions are, then, those associated with trigonometric polynomials m_0 satisfying

$$|m_0(z)|^2 = p_n(\sin^2(z/2)).$$

In order to estimate the regularity index of these scaling functions, we will use Theorem 4.1. In order to compute the λ-limit formed with respect to the polynomial q_n and the transformation W, we need several simple lemmas.

LEMMA 5.1. *We have*

$$q_n(1) = \frac{2n+1}{n+1}\binom{2n}{n}, \qquad q_n'(1) = \frac{n(2n+1)}{n+2}\binom{2n}{n}.$$

Proof. We set $x = 1$ in the differential equation (5.4). Then we differentiate this equation and again set $x = 1$. □

LEMMA 5.2. *We have*

$$\frac{q_n'(x)}{q_n(x)} \geq \frac{q_n'(1)}{q_n(1)} = \frac{n(n+1)}{n+2} \text{ for all } 0 \leq x \leq 1.$$

Proof. Since q_n is a polynomial of degree n with nonnegative coefficients, we obtain from Lemma 5.1 that

$$q_n(x) \geq q_n(1)x^n = \frac{2n+1}{n+1}\binom{2n}{n}x^n \tag{5.5}$$

and

$$q_n'(x) \geq q_n'(1)x^{n-1} = \frac{n(2n+1)}{n+2}\binom{2n}{n}x^{n-1}. \tag{5.6}$$

By (5.4) and (5.6),

$$\begin{aligned} q_n'(x) &= (n+1)q_n(x) + xq_n'(x) - (2n+1)\binom{2n}{n}x^n \\ &\geq (n+1)q_n(x) - \frac{4n+2}{n+2}\binom{2n}{n}x^n. \end{aligned}$$

By (5.5),

$$q_n'(x) \geq (n+1)q_n(x) - \frac{4n+2}{n+2}\frac{n+1}{2n+1}q_n(x) = \frac{n(n+1)}{n+2}q_n(x). \qquad \square$$

LEMMA 5.3. *We have*

$$\frac{q_n'(x)}{q_n(x)} \leq \frac{n}{x} \text{ for all } 0 < x \leq 1.$$

Proof. This inequality holds for all polynomials of degree n with nonnegative coefficients. □

LEMMA 5.4. *We have*

$$q_n(x)q_n(W(x)) \leq q_n(3/4)^2 \text{ for all } 3/4 \leq x \leq 1.$$

Proof. Since this inequality is an equality when $x = 3/4$, it is sufficient to prove that the derivative of $q_n(x)q_n(W(x))$ is nonpositive for all x between $3/4$ and 1. This is equivalent to the inequality

$$-W'(x)\frac{q_n'(W(x))}{q_n(W(x))} \geq \frac{q_n'(x)}{q_n(x)}$$

for the same x. Because of Lemmas 5.2 and 5.3, it is enough to show that

$$(8x-4)\frac{n(n+1)}{n+2} \geq \frac{n}{x}$$

or

$$8x^2 - 4x \geq \frac{n+2}{n+1}$$

for all x between $3/4$ and 1. This is true because

$$8x^2 - 4x \geq \frac{3}{2} \geq \frac{n+2}{n+1}$$

for all $3/4 \leq x \leq 1$ and all positive integers n. □

We can now prove the main result of this section.

THEOREM 5.5. *Let the numbers $\lambda_m, \mu_m, \lambda$ be defined as in §3 with respect to $f = q_n$ and the transformation $T = W$. Then $\lambda = \mu_2 = q_n(3/4)$. In particular, the regularity index σ_n of Daubechies's scaling function associated with the polynomial p_n (which has support width $2n+1$) satisfies*

$$-\log_4 p_n(3/4) - 1/2 \leq \sigma_n \leq -\log_4 p_n(3/4).$$

Proof. If $0 \leq x \leq 3/4$, then $q_n(x) \leq q_n(3/4)$ because q_n is increasing for $x \geq 0$. Together with Lemma 5.4 this shows that

$$\min\{q_n(x), (q_n(x)q_n(W(x)))^{1/2}\} \leq q_n(3/4) \text{ for all } 0 \leq x \leq 1.$$

Since we have equality for $x = 3/4$, this proves that $\lambda \leq \mu_2 = q_n(3/4)$. Now (3.7) shows that $\lambda = \mu_2 = q_n(3/4)$.

The final statement of the theorem follows from Theorem 4.1 because $L = n+1$ and

$$L - \log_4 \lambda = n + 1 - \log_4 q_n(3/4) = -\log_4 p_n(3/4). \quad \square$$

The numbers $p_n(3/4)$ appearing in the previous theorem are given by

$$p_n(3/4) = \alpha_n \int_{3/4}^{1} W(t)^n dt.$$

The asymptotic behavior of this sequence as $n \to \infty$ can be found by standard methods. We only need that

$$\frac{1}{n}\log_4 p_n(3/4) \to \log_4(3/4) \text{ as } n \to \infty, \tag{5.7}$$

which, for example, follows from the following lemma (proof omitted).

LEMMA 5.6. *Let h be a nonnegative continuously differentiable function defined on $[a,b]$. Let $H = \max\{h(x) : a \le x \le b\} > 0$. Then there is a positive constant ϵ (independent of n) such that*

$$\frac{\epsilon}{n} H^n \le \int_a^b h(t)^n dt \le (b-a)H^n.$$

In combination with Theorem 5.5, the relation (5.7) implies the following corollary.

COROLLARY 5.7. *The asymptotic regularity ratio (as defined in the introduction) of the sequence of Daubechies's scaling functions is given by*

$$-\frac{1}{2}\log_4(3/4) = 0.1037\ldots.$$

We remark that the results of this section are known. In particular, Lemma 5.4 and Theorem 5.5 are proved in [2]; see also [5, §7.1]. Lemma 5.4 is contained in [5, Lem. 7.1.8] but the proofs are different. Corollary 5.7 is proved in [13].

6. Sequences of highly regular scaling functions. We now return to the more general sequences of scaling functions introduced in §4. To be more specific, we assume that the polynomial r is given by

$$r(y) = y^M (y-a_1)(y-a_2)\ldots(y-a_K), \tag{6.1}$$

where the numbers a_j satisfy $0 < a_1 < a_2 < \cdots < a_K < 1$. In the previous section, we had $M = 1$ and $K = 0$. The idea is to choose $K > 0$ to obtain sequences of more regular scaling functions. If, for fixed support width $2n(M+K)+1$ of ϕ_n, we increase K, then the number $L = nM+1$ of Theorem 4.1 decreases; however, λ also decreases for appropriately chosen zeros $a_1, \ldots, a_K$.

Let p_n be the polynomial defined by (4.4) for even n. By estimating the integral in (4.4) we obtain

$$p_n(x) \le \frac{1-x}{\alpha_n} R(W(x))^n \text{ for all } 1/2 \le x \le 1,$$

where R is defined by

$$R(y) = \max\{|r(t)| : 0 \le t \le y\}, \qquad 0 \le y \le 1.$$

We also have

$$p_n(x) \le \frac{1-x}{\alpha_n} R(1)^n \text{ for } 0 \le x \le 1/2.$$

Therefore, if we define a continuous function f on $[0,1]$ by

$$f(x) = \begin{cases} (1-x)^{-M} & \text{if} \quad 0 \le x < 1/2, \\ R(W(x))(1-x)^{-M}/R(1) & \text{if} \quad 1/2 \le x < 1, \\ 4^M \prod_{j=1}^K a_j / R(1) & \text{if} \quad x = 1, \end{cases} \tag{6.2}$$

then p_n satisfies

$$p_n(x) \le \frac{1}{\alpha_n}(1-x)^{nM+1} f(x)^n R(1)^n \text{ for } 0 \le x \le 1. \tag{6.3}$$

We now use Theorem 4.1 to obtain the following lower bound for the asymptotic regularity ratio of the sequence ϕ_n.

THEOREM 6.1. *Let r be a polynomial of the form* (6.1). *Then the sequence of orthonormal scaling functions ϕ_n (n even) derived from r has an asymptotic regularity ratio greater than or equal to*

$$\frac{M - \log_4 \lambda}{2(M+K)},$$

where λ is the limit of the sequence λ_m defined in §3 *for the function f of* (6.2) *with respect to the transformation W.*

Proof. The polynomial p_n has a zero of exact order $L = nM + 1$ at $x = 1$. Let q_n be the polynomial defined by

$$p_n(x) = (1-x)^{nM+1} q_n(x).$$

Then (6.3) shows that

$$q_n(x) \le \frac{1}{\alpha_n} f(x)^n R(1)^n \text{ for } 0 \le x \le 1. \tag{6.4}$$

By Lemma 5.6, we find $\epsilon > 0$ such that

$$\alpha_n \ge \frac{\epsilon}{n} R(1)^n. \tag{6.5}$$

Let $\lambda(q_n)$ be the λ-limit of §3 corresponding to $f = q_n$ and $T = W$. Then the inequalities (6.4) and (6.5) show that

$$\lambda(q_n) \le \frac{n}{\epsilon} \lambda^n.$$

By Theorem 4.1, the regularity index σ_n of ϕ_n satisfies

$$\sigma_n \ge nM + 1 - n \log_4 \lambda - \log_4(n/\epsilon) - \frac{1}{2}.$$

The support width of ϕ_n is equal to the degree of p_n which is $2n(M+K)+1$. Therefore, the asymptotic regularity ratio of the sequence ϕ_n is at least

$$\frac{M - \log_4 \lambda}{2(M+K)}.$$

This completes the proof. □

The simplest upper bound for λ is $\lambda_1 = \mu_1$, i.e., the maximum of $f(x)$ for $0 \le x \le 1$. We can write

$$\mu_1 = \frac{\sup\{|r(W(x))|(1-x)^{-M} : 1/2 \le x < 1\}}{\max\{|r(y)| : 0 \le y \le 1\}} \tag{6.6}$$

because the supremum of $\{R(W(x))(1-x)^{-M} : 1/2 \le x < 1\}$ is equal to that of $\{|r(W(y))|(1-y)^{-M} : 1/2 \le y < 1\}$. This follows from the observation that, for every $1/2 \le x < 1$, there exists $x \le y < 1$ such that $R(W(x)) = |r(W(y))|$. Consequently, $R(W(x))(1-x)^{-M} \le |r(W(y))|(1-y)^{-M}$.

Let us consider an example. Let $M = 2$ and $K = 1$. Then r is of the form

$$r(y) = y^2(y - a), \text{ where } 0 < a < 1.$$

We estimate the number λ by μ_1. According to (6.6), we have

$$\mu_1 = \frac{16 \max\{|x^2(W(x) - a)| : 1/2 \le x \le 1\}}{\max\{|y^2(y - a)| : 0 \le y \le 1\}}.$$

The maximum in the numerator is equal to

$$16 \max\{|x_0^2(W(x_0) - a)|, a\},$$

where

$$x_0 = \frac{3 + \sqrt{9 - 8a}}{8}.$$

If we choose $a = 0.2762\ldots$, then $x_0^2|W(x_0) - a| = a$ and $\mu_1 = 16a/(1 - a)$. We now apply Theorem 6.1 and obtain that the asymptotic regularity ratio of the sequence of scaling functions derived from r is at least

$$\frac{2 - \log_4(16a/(1 - a))}{6} = 0.115\ldots.$$

This ratio is greater than that obtained in Corollary 5.7.

This result can be improved if we use $\mu_2, \mu_3, \ldots$ as upper bounds for λ. Computer calculations using μ_3 as an upper bound for λ led to the choices in Table 1 of the polynomial (6.1) and estimates for the asymptotic regularity ratio of the corresponding sequence of scaling functions. By using upper bounds μ_m for larger m it is possible to find asymptotic regularity ratios close to 0.19.

TABLE 1

M	K	$a_1 < \cdots < a_K$	asymp. reg. ratio $\ge$
1	1	0.43	0.147
2	1	0.52	0.155
1	2	$0.09 < 0.56$	0.157
2	2	$0.17 < 0.57$	0.162
3	3	$0.08 < 0.36 < 0.64$	0.171
4	4	$0.05 < 0.26 < 0.45 < 0.67$	0.175

We conclude this paper with the following open questions:

1. Are there more efficient methods for estimating the limit λ of §3?

2. How should we choose the integers M and K and the zeros $a_1, \ldots a_K$ of r in a systematic way to obtain large asymptotic regularity ratios? ($M = K$ seems to be good.)

3. What is the largest asymptotic regularity ratio we can find by the presented method?

4. What is the largest asymptotic regularity ratio for sequences of orthonormal scaling functions?

REFERENCES

[1] A. COHEN, *Construction de bases d'ondelettes α-hölderiennes*, Rev. Mat. Iberoamericana, 6 (1990), pp. 91–108.

[2] A. COHEN AND J. P. CONZE, *Régularité des bases d'ondelettes et mesures ergodiques*, Rev. Mat. Iberoamericana, 8 (1992), pp. 351–365.

[3] I. DAUBECHIES, *Orthonormal bases of compactly supported Wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[4] ———, *Orthonormal bases of compactly supported wavelets, II. Variations on a theme*, SIAM J. Math. Anal., 24 (1993), pp. 499–519.

[5] ———, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[6] I. DAUBECHIES AND J. C. LAGARIAS, *Two-scale difference equations, I. Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[7] ———, *Two-scale difference equations, II. Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[8] T. EIROLA, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[9] P. R. HALMOS, *Lectures on Ergodic Theory*, The Mathematical Society of Japan, Tokyo, 1956.

[10] G. POLYA AND G. SZEGÖ, *Aufgaben und Lehrsätze aus der Analysis*, Springer-Verlag, Berlin, New York, 1964.

[11] O. RIOUL, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal., 23 (1992), pp. 1544–1576.

[12] L. VILLEMOES, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.

[13] H. VOLKMER, *On the regularity of wavelets*, IEEE Trans. Inform. Theory, 38 (1992), pp. 872–876.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1089–1112, September 1995

A STEFAN PROBLEM FOR A REACTION-DIFFUSION SYSTEM*

AVNER FRIEDMAN†, DAVID S. ROSS‡, AND JIANHUA ZHANG§

Abstract. The paper deals with a Stefan problem for a system of three weakly coupled semilinear parabolic equations. The system describes dissolution of a spherical particle in solution. The dissolved species A reacts chemically with species B already in the solution, thereby forming species C. Species C diffuses in the solution and some of it adsorbs to the particle's boundary and gradually shuts down the dissolution. It is shown that the mathematical model has a unique solution with finite shut-down time. When the reaction rate K increases to infinity, the limit model should exhibit phase separation between A and B, and it thus has two free boundaries: the particle's boundary and the $A-B$ interface. It is proved, in the case in which A and B diffuse at the same rate, that the solution with finite K converges to the solution of the limit problem, and the A phase in the limit problem disappears in finite time.

Key words. Stefan problem, reaction-diffusion system, asymptotic estimates, dissolution

AMS subject classifications. 35K57, 35R35, 35B40

1. The model.

Consider a solid spherical particle composed of chemical A with uniform concentration A^*. The particle is in a solution of chemical B. As the particle dissolves, the A that enters the solution reacts with B to form chemical C. Then C diffuses in the solution and some of it reaches the solid particle and adsorbs to its surface. The presence of the adsorbed C inhibits the dissolution and ultimately shuts it down entirely.

Assuming radially symmetric data and radially symmetric functions A, B, C, we denote by $r = R(t)$ the radius of the solid sphere at time t. Then the equations

$$\frac{\partial A}{\partial t} = D_A \Delta A - KAB, \tag{1.1}$$

$$\frac{\partial B}{\partial t} = D_B \Delta B - KAB, \tag{1.2}$$

$$\frac{\partial C}{\partial t} = D_C \Delta C + KAB \tag{1.3}$$

hold in $\{r > R(t)\}$, where K is the reaction rate and D_A, D_B, D_C are the diffusion coefficients. These equations indicate that A and B are lost in a second-order reaction in which C is formed and all three species diffuse. In the standard mass-action model of chemical kinetics, the concentrations are all expressed in moles/liter and the coefficient K, the second-order reaction rate, is expressed in liters/(mole-sec). Then KAB is the number of moles per liter per second that undergoes reaction; in our case, A and B are consumed, C is created, the same number of moles of A and B are lost, and this number of moles of C is created. A nice reference for this material is the book by Erdi and Toth [4].

* Received by the editors September 20, 1993; accepted for publication (in revised form) February 10, 1994.

† Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455. The research of this author was partially supported by National Science Foundation grant DMS87-22187.

‡ Applied Mathematics and Statistics Group, Computational Science Laboratory, Eastman Kodak Company, Rochester, New York 14650-2205.

§ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

Next,

$$\frac{dR}{dt} = \alpha \frac{\partial A}{\partial r} \quad \text{on} \quad r = R(t), \tag{1.4}$$

where α is a positive constant, i.e., the rate at which the radius of the particle decreases is proportional to the flux of species A away from the particle. We also have

$$\frac{\partial B}{\partial r} = 0 \quad \text{on} \quad r = R(t), \tag{1.5}$$

i.e., there is no flux of B through the particle's surface and B does not undergo any surface reaction.

The adsorption of C to the surface is proportional to the local saturation; it is given by an empirical law $D_C \partial C/\partial r = \gamma C^n$ for some positive constants γ, n (see [13, pp. 104–105]); for definiteness we take $n = 4$, i.e.,

$$D_C \frac{\partial C}{\partial r} = \gamma C^4. \tag{1.6}$$

However, all the results of this paper remain valid with minor changes if we replace γC^4 by any other monotone increasing function $f(C)$ with $f(0) = 0$, $f(C) > 0$ for $C > 0$.

The boundary conditions at $r = \infty$ are

$$A(\infty, t) = 0, \quad B(\infty, t) = B^*, \quad C(\infty, t) = 0, \tag{1.7}$$

where B^* is a positive constant.

We now impose initial conditions. First,

$$R(0) = R_0 > 0. \tag{1.8}$$

Next we assume that

$$A(r, 0) = A_0(r), \quad B(r, 0) = B_0(r), \quad C(r, 0) = C_0(r) \tag{1.9}$$

for $r > R_0$, where A_0, B_0, C_0 are approximately $0, B^*, 0$ (i.e., initially, only B is mostly present in the solution and its concentration is nearly uniform). We also assume that the initial conditions are smooth and fit smoothly with the boundary conditions

$$\begin{cases} A_0 \in C^{2+\nu}[R_0, \infty), \ A_0'(r) \le 0, \ A_0(R_0) = A^*, \\ A_0(r) = 0 \quad \text{if } r > R_0 + \delta_1 \text{ for some } \delta_1 > 0; \end{cases} \tag{1.10}$$

$$\begin{cases} B_0 \in C^{2+\nu}[R_0, \infty), \ B_0(r) \ge 0, \ B_0'(r) \ge 0, \\ B_0(r) = B^* \text{ if } r \ge R_0 + \delta_2 \quad \text{for some} \quad \delta_2 > 0; \end{cases} \tag{1.11}$$

$$\begin{cases} C_0(r) \in C^2[R_0, \infty), \ C_0(r) \ge 0, \\ C_0(r) = 0 \quad \text{if } r > R_0 + \delta_3 \text{ for some } \delta_3 > 0, \\ D_C \dfrac{\partial C_0}{\partial r} = \gamma C_0^4 \quad \text{at} \quad r = R_0 \end{cases} \tag{1.12}$$

for some $0 < \nu < 1$.

Finally we need to determine the boundary condition for A at the particle's surface. The flux of A from the particle depends on the amount of C that is adsorbed to the surface. On a portion of $\{r = R\}$ where there is no adsorbed C, $A = A^*$, the saturation concentration of A; local thermodynamic equilibrium is established instantaneously. On a portion which is fully covered by $C, \partial A/\partial r = 0$, i.e., the dissolution shuts down. This is actually a microscopic statement, which we shall now "average." We shall use the "weighted average"

$$\zeta(t) = \frac{\beta \int_0^t R^2(s) D_C \frac{\partial C}{\partial r}\Big|_{r=R(s)} ds + \delta}{R^2(t)}, \tag{1.13}$$

where β is a positive empirical parameter and δ is a small positive parameter such that $\delta/R_0^2 < 1$. We then impose the boundary condition

$$-\zeta_0(t) D_A \frac{\partial A}{\partial r} + (1 - \zeta_0(t))(A - A^*) = 0 \quad \text{on} \quad r = R(t), \tag{1.14}$$

where

$$\zeta_0(t) = \min\{\zeta(t), 1\}. \tag{1.15}$$

Thus the dissolution shuts down as soon as $\zeta(t)$ becomes equal to 1. This boundary condition has the basic properties demanded by the physical problem; it reduces to the Dirichlet condition in the absence of adsorbed C, it reduces to the Neumann condition when the surface is covered, and it makes a continuous monotone transition between these two conditions as a function of the fraction of surface area that is covered.

Remark 1.1. The parameter δ in (1.13) ensures that $\zeta(0) > 0$ and, therefore, the boundary condition in (1.14) does not degenerate at $t = 0$. All the results of this paper, however, except uniqueness, extend to the case $\delta = 0$ by simply going to the limit with $\delta \to 0$. If $\delta = 0$, the solution is not smooth at $(R_0, 0)$ and our proof of uniqueness (for the case $\delta > 0$) does not carry through.

For additional information on the model see [9, Chap. 18].

2. The main results.

DEFINITION 2.1. *We refer to the system* (1.1)–(1.15) *as problem* (P). *By a solution to problem* (P) *we mean* (A, B, C, R) *satisfying* (1.1)–(1.15) *in the classical case; in particular,* $R(t)$ *is continuously differentiable for all* $t \geq 0$.

DEFINITION 2.2. *Suppose* T^* *is such that*

$$\zeta(t) < 1 \quad \text{if} \quad t < T^*, \quad \zeta(t) \geq 1 \quad \text{if} \quad t \geq T^*.$$

Then we call T^* *the* shut-down time.

Note that (1.4) and (1.14) reduce to

$$R(t) = R(T^*), \quad A_r(R(T^*), t) = 0 \quad \text{for} \quad t > T^*. \tag{2.1}$$

THEOREM 2.1. *There exists a unique solution of* (P), *and it has the following properties:* (i) $R(t) > 0$, $R'(t) \leq 0$, *and* $\zeta'(t) > 0$ *for all* $t > 0$; (ii) *it has a finite shut-down time* T^* *and* $R'(t) < 0$ *if* $0 < t < T^*$; (iii) R *and* ζ *belong to* $C^{1+\mu}[0, \infty) \cap C^\infty(0, T^*)$ *for any* $0 < \mu < 1$.

We are interested in the case of fast reaction, i.e., $K \gg 1$. This motivates the study of the solution

$$(A_K,\ B_K,\ C_K,\ R_K) \quad \text{of} \quad (P)$$

as $K \to \infty$. It can be shown that

$$\int_0^T \int_{R_K(t)}^{\infty} A_K B_K r^2 \, dr \, dt \le \frac{\text{const.}}{K}$$

and

$$\frac{\partial}{\partial r} A_K \le 0, \qquad \frac{\partial}{\partial r} B_K \ge 0.$$

It follows, formally, that the limits

$$\begin{aligned} &A = \lim A_K, \quad B = \lim B_K, \quad C = \lim C_K, \\ &R = \lim R_K, \quad \text{and} \quad \zeta = \lim \zeta_K \end{aligned} \tag{2.2}$$

are such that $AB = 0$, i.e.,

$$\begin{aligned} A(r,t) &> 0 \quad \text{if} \quad R(t) < r < S(t), \\ &= 0 \quad \text{if} \quad r > S(t), \end{aligned} \tag{2.3}$$

$$\begin{aligned} B(r,t) &> 0 \quad \text{if} \quad r > S(t), \\ &= 0 \quad \text{if} \quad R(t) < r < S(t) \end{aligned} \tag{2.4}$$

for some function $S(t)$; furthermore,

$$\frac{\partial A}{\partial t} = D_A \Delta A \quad \text{if} \quad R(t) < r < S(t), \quad t > 0, \tag{2.5}$$

$$\frac{\partial B}{\partial t} = D_B \Delta B \quad \text{if} \quad S(t) < r < \infty, \quad t > 0, \tag{2.6}$$

$$\frac{\partial C}{\partial t} = D_C \Delta C - \left(D_A \frac{\partial A}{\partial r}\bigg|_{r=S(t)} \right) \delta(r - S(t)) \quad \text{if} \quad R(t) < r < \infty, \ t > 0, \tag{2.7}$$

$$\frac{dR}{dt} = \alpha \frac{\partial A}{\partial r}\bigg|_{r=R(t)}, \tag{2.8}$$

$$-\zeta_0 D_A \frac{\partial A}{\partial r} + (1 - \zeta_0)(A - A^*) = 0 \quad \text{at} \quad r = R(t), \tag{2.9}$$

where ζ_0 is defined by (1.13), (1.15), and

$$D_C \frac{\partial C}{\partial r} = \gamma C^4 \quad \text{on} \quad r = R(t), \tag{2.10}$$

$$A(S(t), t) = 0, \tag{2.11}$$

$$B(S(t), t) = 0, \tag{2.12}$$

$$D_A \frac{\partial A}{\partial r} = -D_B \frac{\partial B}{\partial r} \quad \text{on} \quad r = S(t). \tag{2.13}$$

Since $A + B \to C$, the generation of C which occurs at $r = S(t)$ is at the same rate as $D_A \partial A/\partial r$ or $-D_B \partial B/\partial r$, which explains both (2.13) and the source term in (2.7). Finally, we have the initial conditions

$$\begin{aligned} A(r,0) &= A_0(r) \quad \text{if} \quad R(0) < r < S(0), \\ B(r,0) &= B_0(r) \quad \text{if} \quad S(0) < r < \infty, \\ C(r,0) &= C_0(r) \quad \text{if} \quad R(0) < r < \infty, \end{aligned} \tag{2.14}$$

$$S(0) = r_0, \quad \text{where } r_0 \text{ is such that } A_0(r_0) = B_0(r_0) \tag{2.15}$$

(r_0 is uniquely determined), and the conditions at $r = \infty$,

$$B(\infty, t) = B^*, \quad C(\infty, t) = 0. \tag{2.16}$$

DEFINITION 2.3. *We shall denote the problem* (P) *by* (P_K) *and refer to the system* (2.5)–(2.16) *as problem* (P_∞). *By a solution to problem* (P_∞) *we mean* (A, B, C, R, S, T_f) *such that all the equations are satisfied in the classical case for* $0 < t < T_f$, *and*

$$\begin{aligned} R(t) < S(t) < \infty \quad &\text{if} \quad t < T_f, \\ S(t) - R(t) \to 0 \quad &\text{if} \quad t \to T_f; \end{aligned}$$

in particular, $R(t)$ *and* $S(t)$ *are continuously differentiable for* $0 < t < T_f$ *and continuous for* $0 \le t \le T_f$ *and* $A_r(r,t)$ *is continuous for* $R(t) \le r \le S(t)$.

The curve $r = S(t)$ is the interface between the separated phases A and B, and T_f is the *final time*, i.e., the time at which phase A has totally disappeared.

THEOREM 2.2. *Assume that* $D_A = D_B$ *and* $C_0(r) \not\equiv 0$. *Then there exists a unique solution to problem* (P_∞) *and it has the following properties:* (i) $R(t) > 0$, $R'(t) \le 0$, *and* $\zeta'(t) > 0$ *for all* $0 < t < T_f$, *and* $R'(t) < 0$ *as long as* $\zeta(t) < 1$; (ii) $T_f < \infty$; (iii) $R, S,$ *and* ζ *belong to* $C^0[0, T_f] \cap C^{1+\mu}[0, T_f) \cap C^\infty(0, T_f)$.

THEOREM 2.3. *Assume that* $D_A = D_B$. *Then, as* $K \to \infty$, *the limits in* (2.2) *exist, where* (A, B, C, R, S, T_f) *is the solution to problem* (P_∞); *the convergence of* A_K, B_K, C_K *is uniform in any compact subset of*

$$\{R(t) < r < S(t),\ 0 \le t < T_f\} \cup \{S(t) < r < \infty,\ 0 \le t < T_f\}, \tag{2.17}$$

and the convergence of R_K *and* ζ_K *is uniform for* $0 \le t < T_f$.

Theorem 2.1 is proved in §§3–6 and Theorems 2.2 and 2.3 are proved in §§7–10.

Remark 2.1. Reaction-diffusion systems of the form

$$A_t = D_A \Delta A - KAB, \quad B_t = D_B \Delta B - KAB \tag{2.18}$$

with $K \to \infty$ have been studied in [1] and [12]. Evans [5] considered (2.18) in a fixed cylinder $\Omega \times (0, T)$ under the assumptions

$$\frac{\partial A}{\partial n} = 0 \quad \text{on} \quad \partial\Omega, \quad \frac{\partial B}{\partial n} = 0 \quad \text{on} \quad \partial\Omega, \quad A_0 B_0 = 0, \tag{2.19}$$

where A_0, B_0 are the nonnegative initial values for A and B, respectively. He proved that, as $K \to \infty, A \to u^+$ and $B \to u^-$, where u is the solution of

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

with $\partial u/\partial n = 0$ on $\partial\Omega$ and $u|_{t=0} = A_0 - B_0$; here

$$a(u) = D_A \quad \text{if} \quad u > 0, \quad a(u) = D_B \quad \text{if} \quad u < 0.$$

(Uniqueness for the limit problem was established in [2].) His proof relies heavily on the assumptions in (2.19) and does not seem to extend to the present case where we have a moving boundary $r = R(t)$ and different boundary conditions from those in (2.19).

Remark 2.2. For the general study of the Stefan problem in n dimensions we refer to [8] and [10] and the references therein. In the case of radially symmetric solutions, existence, uniqueness, and regularity have been established by several methods (see [6] and [7] and the references therein). In the standard Stefan problem one assumes that A vanishes on the free boundary. The condition (1.14) is called a "kinetic" condition. A Stefan problem for one heat equation in one dimension with kinetic condition was studied by Visintin [14] and Xie [15].

3. Local existence for (P). In this section we prove the following theorem.

THEOREM 3.1. *There exists a solution* (A, B, C, R) *of problem* (P) *for* $0 < t < T$, *where* T *is some small positive number.*

The proof is based on a fixed-point argument. For any $N_0 > 0$, set

$$K_R = \{R(t) \in C^{0,1}[0,T], R(0) = R_0, \ -M \le \dot{R}(t) \le 0 \text{ a.e.}\},$$
$$K_\zeta = \left\{\zeta(t) \in C^{0,1}[0,T], \frac{\delta}{R_0^2} \le \zeta(t) \le N_0, \ 0 \le \dot{\zeta}(t) \le N \text{ a.e.}\right\},$$

where

$$M = \frac{\alpha}{D_A}\,\frac{A^* R_0^2}{\delta}$$

and N is a positive constant to be determined. We endow K_R and K_ζ with the $C^0[0,T]$ norm; then $K_R \times K_\zeta$ is a compact set in $C^0[0,T] \times C^0[0,T]$.

For each $(R(t), \zeta(t)) \in K_R \times K_\zeta$ there exists a unique solution $(A(r,t), B(r,t))$ of (1.1), (1.2), (1.5), (1.14), (1.7) with the initial conditions as in (1.9); since the parabolic system is weakly coupled, such a solution exists for any given time T. By the maximum principle,

$$0 \le A(r,t) \le A^*, \quad 0 \le B(r,t) \le B^*. \tag{3.1}$$

Next we prove that

$$A_r(r,t) \le 0, \quad B_r(r,t) \le 0. \tag{3.2}$$

If we differentiate (1.1), (1.2) with respect to r, we get a coupled system of parabolic equations

$$\frac{\partial}{\partial t}A_r - \mathcal{L}A_r = -KAB_r,$$
$$\frac{\partial}{\partial t}B_r - \mathcal{L}B_r = -KBA_r,$$

where $\mathcal{L}$ is an elliptic operator. On $t = 0$ and the boundary $r = R(t)$ we have $A_r \leq 0$, $B_r \geq 0$. We approximate A_r, B_r by solutions $A_r^\varepsilon, B_r^\varepsilon$ satisfying the same parabolic system with initial and boundary conditions given by

$$A_r^\varepsilon = A_r - \varepsilon, \quad B_r^\varepsilon = B_r + \varepsilon.$$

Then $A_r^\varepsilon < 0$, $B_r^\varepsilon > 0$ for $R(t) \leq r < \infty, 0 \leq t \leq T$. Indeed, otherwise there is a smallest t_0 such that $A_r^\varepsilon \leq 0$, $B_r^\varepsilon \geq 0$ for $R(t) \leq r < \infty$, $0 \leq t \leq t_0$, and $A_r^\varepsilon = 0$ or $B_r^\varepsilon = 0$ at some point (r_0, t_0). This is a contradiction of the strong maximum principle applied to A_r^ε or B_r^ε.

If we now let $\varepsilon \to 0$, we obtain assertion (3.2).

Motivated by (1.4), (1.14), we now define

$$\overline{R}(t) = R_0 + \frac{\alpha}{D_A} \int_0^t \frac{1 - \zeta_0(s)}{\zeta_0(s)} [A(R(s), s) - A^*] ds. \tag{3.3}$$

Next we consider the parabolic equation (1.3) in $\overline{R}(t) < r < \infty$, $0 < t < T$ with boundary conditions (1.6) and $C(\infty, t) = 0$ and initial conditions $C(r, 0) = C_0(r)$. Since the functions $KA^*B^*t + \widehat{C}$ $(\widehat{C} = \sup C_0)$ and 0 are a supersolution and subsolution, respectively, the existence of a solution can be established by a fixed-point argument (cf. [6, Chap. 7, §5]). Uniqueness follows by a comparison principle [6].

We now define

$$\overline{\zeta}(t) = \frac{\beta\gamma \int_0^t \overline{R}^2(s) C^4(\overline{R}(s), s) + \delta}{\overline{R}^2(t)} \tag{3.4}$$

and consider the mapping W:

$$W(R(t),\ \zeta(t)) = (\overline{R}(t),\ \overline{\zeta}(t)).$$

If we show that W has a fixed point in $K_R \times K_\zeta$, then this yields a solution to problem (P).

LEMMA 3.2. *W maps $K_R \times K_\zeta$ into itself.*

Proof. From (3.3), (3.4) we get

$$\frac{d}{dt} \overline{R}(t) = \frac{\alpha}{D_A} \frac{1 - \zeta_0(t)}{\zeta_0(t)} (A - A^*), \tag{3.5}$$

$$\frac{d}{dt} \overline{\zeta}(t) = \beta\gamma C^4(\overline{R}(t), t) - \frac{2}{\overline{R}(t)} \overline{\zeta}(t) \frac{d\overline{R}(t)}{dt}. \tag{3.6}$$

From (3.5) we see that

$$-M \leq \frac{d\overline{R}(t)}{dt} \leq 0. \tag{3.7}$$

Since $C \leq KA^*B^*t + \widehat{C}$, we find from (3.4) that $\overline{\zeta}(t) \leq N_0$ if T is small enough. From (3.6) and (3.7) it then follows that

$$0 < \frac{d}{dt} \overline{\zeta}(t) \leq \widehat{N},$$

provided $\widehat{N}$ is a sufficiently large number independent of N (see the definition of K_ζ). Hence, choosing $N = \widehat{N}$ we see that $(\overline{R}, \overline{\zeta})$ belongs to $K_R \times K_\zeta$. □

LEMMA 3.3. *W is continuous (when $K_R \times K_\zeta$ is endowed with the $C^0[0,T] \times C^0[0,T]$ norm).*

Proof. If we use the transformation $\widehat{r} = r - R(t)$ in order to flatten the boundary $r = R(t)$, we get a new parabolic equation for A, where a new term $\dot{R}A_r$ is added to the heat operator. On the lateral boundary $\widehat{r} = 0$,

$$-A_r + a(t)A = b(t),$$

where a, b are Lipschitz continuous and their Lipschitz constants are bounded independently of (R, ζ) in $K_R \times K_\zeta$. We can therefore apply L^p estimates [10] to deduce that

$$\|A_r\|_{L^p(\Omega_T)} + \|A_{rr}\|_{L^p(\Omega_T)} + \|A_t\|_{L^p(\Omega_T)} \le C_1 \tag{3.8}$$

for any $1 < p < \infty$, where

$$\Omega_T = \{R(t) \le r < \infty,\ 0 \le t \le T\}$$

and C_1 is a constant depending on p but not on (R, ζ) and T. By Sobolev's imbedding [11] we then have the Hölder estimate

$$\|A_r\|_{C^\mu(\Omega_T)} \le C_2 \tag{3.9}$$

for some $0 < \mu < 1$ and C_2 independent of T and (R, ζ).

We now proceed to prove that W is continuous.

Suppose (R_n, ζ_n) and (R, ζ) belong to $K_R \times K_\zeta$ and $R_n \to R$, $\zeta_n \to \zeta$ in the $C^0[0,T]$ norm. We need to prove that

$$W(R_n, \zeta_n) \to W(R, \zeta).$$

Define $A_n, B_n, \overline{\zeta}_n, \overline{R}_n$, and C_n corresponding to R_n, ζ_n, so that $W(R_n, \zeta_n) = (\overline{R}_n, \overline{\zeta}_n)$. Applying the estimates (3.8), (3.9) to A_n and similar estimates to B_n and C_n, we can easily show that any subsequence of n's has a subsequence for which

$$A_n \to \widetilde{A}, \quad B_n \to \widetilde{B},$$

where $\widetilde{A}, \widetilde{B}$ satisfy the same parabolic system in $R(t) < r < \infty$, $0 < t < T$ which A, B satisfy. By uniqueness it follows that $\widetilde{A} = A$ and $\widetilde{B} = B$. Therefore

$$A_n \to A,\ B_n \to B \quad \text{uniformly in } \rho_n(t) \le r < \infty,\ 0 \le t \le T, \tag{3.10}$$

where $\rho_n(t) = \max\{R_n(t),\ R(t)\}$.

We proceed to prove that

$$\overline{R}_n(t) - \overline{R}(t) = \frac{\alpha}{D_A}\left\{ \int_0^t \frac{1 - \zeta_{n,0}(s)}{\zeta_{n,0}(s)}[A_n(R_n(s), s) - A^*]ds \right.$$

$$\left. - \int_0^t \frac{1 - \zeta_0(s)}{\zeta_0(s)}[A(R(s), s) - A^*]ds \right\} \to 0, \tag{3.11}$$

where $\zeta_{n,0} = \min\{\zeta_n, 1\}$. By (3.9),

$$|A_n(R_n(s), s) - A_n(\rho_n(s), s)| \le C_1 |R_n(s) - R(s)|^{\mu},$$
$$|A(R(s), s) - A(\rho_n(s), s)| \le C_1 |R_n(s) - R(s)|^{\mu}.$$

Also using (3.10) we conclude that

$$\left| \frac{\alpha}{D_A} \int_0^t \frac{1 - \zeta_{n,0}(s)}{\zeta_{n,0}(s)} [A_n(R_n(s), s) - A(R(s), s)] ds \right| \le C_2 \int_0^t |R_n(s) - R(s)|^{\mu} ds + \varepsilon_n,$$

where $\varepsilon_n \to 0$ uniformly in t as $n \to \infty$. Since $\zeta_n(t) \to \zeta(t)$ uniformly, we deduce from the expression for $\overline{R}_n - \overline{R}$ in (3.11) that

$$|\overline{R}_n(t) - \overline{R}(t)| \to 0 \quad \text{uniformly in } \ t \in [0, T].$$

Similarly, we can prove that $C_n \to C$ for $\overline{R}(t) < r < \infty$, $0 \le t \le T$, and $\overline{\zeta}_n(t) \to \overline{\zeta}(t)$ uniformly in $0 \le t \le T$, and this completes the proof of Lemma 3.3.

Proof of Theorem 3.1. From Lemmas 3.2 and 3.3 we have that W maps the compact set $K_R \times K_\zeta$ into itself and is continuous. By the Schauder fixed-point theorem W has a fixed point (R, ζ), and this completes the proof of Theorem 3.1. □

4. Uniqueness. In this section we prove the following theorem.

THEOREM 4.1. *For any $T > 0$, problem (P) has at most one solution.*

Proof. Suppose there are two solutions $(A_i, B_i, C_i, R_i, \zeta_i)$ $(i = 1, 2)$. Set

$$\widehat{A}_i(x,t) = A_i(r,t), \quad \widehat{B}_i(x,t) = B_i(r,t), \quad \widehat{C}_i(x,t) = C_i(r,t),$$

where $x = r - R_i(t)$. Then

(4.1)

$$\widehat{A}_{i,t} = D_A \left(\widehat{A}_{i,xx} + \frac{2}{x + R_i} \widehat{A}_{i,x} \right) + \dot{R}_i \widehat{A}_{i,x} - K \widehat{A}_i \widehat{B}_i \quad \text{in} \quad Q_T = \{x > 0,\ t > 0\},$$
$$-D_A \zeta_{i,0} \widehat{A}_{i,x} + (1 - \zeta_{i,0})(\widehat{A}_i - A^*) = 0 \quad \text{for} \quad x = 0,\ t > 0,$$
$$\widehat{A}_i(x, 0) = A_0(x + R_0) \quad \text{for} \quad x > 0,$$
$$\lim_{x \to \infty} \widehat{A}_i(x,t) = 0 \quad \text{for} \quad t > 0.$$

Similar systems can be written for $\widehat{B}_i$ and $\widehat{C}_i$. Set

$$u = \widehat{A}_1 - \widehat{A}_2, \quad v = \widehat{B}_1 - \widehat{B}_2, \quad w = \widehat{C}_1 - \widehat{C}_2$$

and

$$\rho = R_1 - R_2, \quad \eta = \zeta_1 - \zeta_2.$$

Then

$$\dot{\rho}(t) = \frac{\alpha}{D_A} \left[\frac{1 - \zeta_{1,0}(t)}{\zeta_{1,0}(t)} (\widehat{A}_1(x,0) - A^*) - \frac{1 - \zeta_{2,0}(t)}{\zeta_{2,0}(t)} (\widehat{A}_2(x,0) - A^*) \right],$$

so that

$$|\dot{\rho}(t)| \le C[|u(0,t)| + |\eta(t)|]. \tag{4.2}$$

Next, from (1.13) we easily estimate

$$|\eta(t)| \le C\left[\|\rho\|_{L^\infty(0,T)} + \|w(0,t)\|_{L^\infty(0,T)}\right]. \tag{4.3}$$

Since $\rho(0) = 0$, (4.2) yields

$$\|\rho\|_{L^\infty(0,T)} \le CT\left[\|u(0,t)\|_{L^\infty(0,T)} + \|\eta\|_{L^\infty(0,T)}\right]. \tag{4.4}$$

Substituting this into (4.3), we get

$$\|\eta\|_{L^\infty(0,T)} \le \frac{CT}{1-CT}\left[\|u(0,t)\|_{L^\infty(0,T)} + \|w(0,t)\|_{L^\infty(0,T)}\right] \tag{4.5}$$

if T is such that $CT < 1$. Finally, using (4.5) in (4.4) we get

$$\|\rho\|_{L^\infty(0,T)} \le CT\left[\|u(0,t)\|_{L^\infty(0,T)} + \|w(0,t)\|_{L^\infty(0,T)}\right]. \tag{4.6}$$

From (4.1) we see that u satisfies a parabolic system

$$\begin{aligned}
&u_t = D_A u_{xx} + \left(\frac{2D_A}{x+R_1} + \dot{R}_1\right) u_x - K\widehat{B}_1 u + F \quad \text{in} \quad Q_T,\\
&-D_A\zeta_{1,0}u_x + (1-\zeta_{1,0})u = G \quad \text{if} \quad x = 0,\ t > 0,\\
&u(x,0) = 0, \quad x > 0,\\
&\lim_{x\to\infty} u(x,t) = 0, \quad t > 0,
\end{aligned} \tag{4.7}$$

where

$$\begin{aligned}
|F| &\le C\left[\|\dot{\rho}\|_{L^\infty(0,T)} + \|v\|_{L^\infty(Q_T)}\right]\\
&\le C\left[\|u\|_{L^\infty(Q_T)} + \|v\|_{L^\infty(Q_T)} + \|w\|_{L^\infty(Q_T)}\right]
\end{aligned}$$

by (4.2), (4.5), and

$$|G| \le C|\eta(t)| \le CT\left[\|u\|_{L^\infty(Q_T)} + \|w\|_{L^\infty(Q_T)}\right].$$

It is easy to see that the function

$$C(t+T)\cdot\left[\|u\|_{L^\infty(Q_T)} + \|v\|_{L^\infty(Q_T)} + \|w\|_{L^\infty(Q_T)}\right]$$

is a supersolution to (4.7), so that

$$\|u\|_{L^\infty(Q_T)} \le CT\left[\|u\|_{L^\infty(Q_T)} + \|v\|_{L^\infty(Q_T)} + \|w\|_{L^\infty(Q_T)}\right].$$

The same estimate can similarly be established for v and w. Hence if T is small enough then $u = v = w = 0$ in Q_T. We can now proceed step-by-step to prove that $u = v = w = 0$ for all t, as long as the two solutions exist.

5. Global existence. In this section we prove that there exists a solution to problem (P) for all time. We first recall that, as long as $\zeta(t) \leq 1$,

$$\dot{R}(t) = \frac{\alpha}{D_A} \frac{1-\zeta(t)}{\zeta(t)}(A - A^*), \tag{5.1}$$

$$\dot{\zeta}(t) = \beta\gamma C^4(R(t), t) - \frac{2\zeta(t)}{R(t)} \dot{R}(t). \tag{5.2}$$

Hence

$$\dot{R}(t) < 0, \quad \dot{\zeta}(t) > 0. \tag{5.3}$$

LEMMA 5.1. *There exists a positive constant R_* such that*

$$R(t) \geq R_* \quad \text{as long as} \quad \zeta(t) \leq 1. \tag{5.4}$$

Proof. Take λ and ε_1 positive and small such that

$$C^4(R(t), t) \leq \varepsilon_1 \quad \text{and} \quad \frac{1}{2} R_0 < R(t)$$

for all $\lambda \leq t \leq 2\lambda$. For $t > 2\lambda$,

$$\zeta(t) \geq \frac{\beta\gamma\varepsilon_1\lambda + \delta}{R^2(t)}.$$

Hence $R(t) > R_*$ as long as $\zeta(t) \leq 1$, where $R_* = (\beta\gamma\varepsilon_1\lambda)^{1/2}$. □

Remark 5.1. Note that the lower bound R_* is independent of the regularizing parameter δ; cf. Remark 1.1

THEOREM 5.2. *There exists a global solution to problem (P), and $R \in C^{1+\mu}[0,\infty) \cap C^\infty(0, T^*)$ for any $0 < \mu < 1$.*

Proof. Suppose we already have a solution for $0 \leq t \leq T$, where T is a positive number not necessarily small. By Lemma 5.1, $R(t) \geq R_*$ for $0 \leq t \leq T$.

Since $R(t) \geq R_* > 0$ (R_* independent of t), a review of the proof of local existence shows that the solution can be extended to $0 \leq t \leq T + T_0$ provided T_0 is a small positive constant depending only on an a priori bound on $\sup |\dot{R}(t)|$. By (5.1), $|\dot{R}(t)|$ is indeed uniformly bounded (independently on t) and, therefore, the solution to problem (P) can be extended step-by-step to all $t > 0$.

To prove the a priori regularity of R and ζ, we perform a change of variables

$$\widehat{A}(x,t) = A(r,t), \quad \widehat{B}(x,t) = B(r,t), \quad \widehat{C}(x,t) = C(r,t),$$

where $x = r - R(t)$. Then, for any $T > 0$, $\widehat{A}$ satisfies

$$\begin{aligned}
&\widehat{A}_t = D_A\left(\widehat{A}_{xx} + \frac{2}{x+R}\widehat{A}_x\right) + \dot{R}A_x - K\widehat{A}\widehat{B} \quad \text{in} \quad Q_T,\\
&-D_A\zeta_0\widehat{A}_x + (1-\zeta_0)(\widehat{A} - A^*) = 0 \quad \text{for} \quad x = 0, \quad 0 < t < T,\\
&\widehat{A}(x,0) = A_0(x + R_0), \quad x > 0,\\
&\lim_{x\to\infty} \widehat{A}(x,t) = 0, \quad 0 < t < T.
\end{aligned} \tag{5.5}$$

Since $\dot{R}(t)$ is uniformly bounded, the same is true of $\dot{\zeta}(t)$ by (5.2) (recall that $C(R(t),t)$ is bounded by $KA^*B^*T+\widehat{C}$). We can therefore apply the L^p parabolic estimates [10] to $\widehat{A}$ and conclude that

$$\int_{Q_T} (|\widehat{A}_r| + |\widehat{A}_{rr}| + |\widehat{A}_t|)^p \le C_{p,T}$$

for any $p > 1$. This implies that

$$\widehat{A} \in C_{r,t}^{\mu,\mu/2}(Q_T) \quad \text{for} \quad 0 < \mu < 1$$

and yields the $C^{1+\mu/2}[0,T]$ regularity of $R(t)$.

Similar L^p estimates can be established for $\widehat{B}$ and $\widehat{C}$ and then, from (5.2), one deduces the $C^{1+\mu/2}[0,T]$ regularity of $\zeta(t)$.

The above arguments can be used to prove step-by-step the $C^{1+\mu}[0,T]$ and $C^\infty(0,T)$ regularity of $R(t)$ and $\zeta(t)$ for any $0 < T \le T^*$. □

6. Finite shut-down.

LEMMA 6.1. *If $\zeta(t) < 1$ for all $t > 0$ then*

$$\lim_{t\to\infty} \dot{R}(t) = 0. \tag{6.1}$$

Proof. By the interior regularity argument used in the proof of Theorem 5.2, if $\zeta(t) < 1$ for all $t > 0$ then

$$|\dot{R}|_{C^\alpha(T,T+1)} \le C \quad \forall\, T > 1, \tag{6.2}$$

where C is a constant independent of T. Now suppose that (6.1) is not true. Then there exists a constant $\mu > 0$ and a sequence $t_n \to \infty$ such that $t_{n+1} > t_n + 1$ and

$$\dot{R}(t_n) \le -2\mu$$

for all n. Then, using (6.2) we also have

$$\dot{R}(t) < -\mu \quad \text{if} \quad |t - t_n| < \lambda$$

for some $\lambda > 0$ independent of n. But then, by the mean value theorem,

$$R(t_n + \lambda) - R(t_n - \lambda) = 2\lambda \dot{R}(\tilde{t}_n) < -2\lambda\mu$$

for all $n \ge 1$, which is impossible since $R(t)$ is monotone decreasing and positive for all $t > 0$. □

LEMMA 6.2. *If $\zeta(t) < 1$ for all $t > 0$ then there exists a positive constant M such that*

$$C_r(R(t),t) \ge \frac{M}{t^6} \tag{6.3}$$

for all t large.

Proof. For any $t_0 > 0$, $C(r,t_0) > 0$ for all $r \ge R(t_0)$. Choose $R(t_0) < a < b < \infty$ such that

$$C(r,t_0) > \varepsilon_0 \chi_{[a,b]}(r) \equiv c_0(r) \tag{6.4}$$

for some $\varepsilon_0 > 0$ and define

$$\widetilde{C}(r,t) = \int_{\mathbb{R}^3} \frac{e^{-\frac{|x-y|^2}{4\sqrt{D_C}\,(t-t_0)}}}{(t-t_0)^{3/2}}\, \varepsilon c_0(|y|)dy, \quad 0 < \varepsilon < 1 \tag{6.5}$$

for $r = |x| > R(t),\ t > t_0$. Since $\widetilde{C}(R(t_0), t_0) = 0$,

$$C(R(t),t) > \widetilde{C}(R(t),t) \quad \text{if} \quad t_0 \le t \le t_1$$

for some $t_1 > t_0$. On the other hand, for $r = R(t),\ t \ge t_1$ we have

$$-D_C\widetilde{C}_r + \gamma\widetilde{C}^4 \le -D_C\varepsilon \int_{\{a<|y|<b\}} \frac{|y| - R(t)}{(4\pi)^{3/2}2\sqrt{D_C}\,(t-t_0)^{5/2}}\, e^{-\frac{|R(t)-y|^2}{4\sqrt{D_C}\,(t-t_0)}} + \frac{C_1\varepsilon^4}{(t-t_0)^6} < 0$$

if ε is sufficiently small. Also recalling (6.4), (6.5) and noting that $\widetilde{C}$ is a subsolution of (1.3), we conclude, by comparison, that if ε is small enough then

$$C(r,t) > \widetilde{C}(r,t) \quad \text{if} \quad t > t_0,$$

from which (6.3) follows upon using (1.6). □

We finally prove the following theorem.

THEOREM 6.3. *There is a finite shut-down time T^*.*

Proof. Suppose $\zeta(t) < 1$ for all $t > 0$. For $r = R(t)$ we can write

$$\dot{\zeta}(t) = \beta D_C C_r - \zeta(t)\frac{2\dot{R}(t)}{R(t)} = \beta D_C C_r - \frac{2\alpha}{R(t)}\,\zeta(t)A_r$$
$$= \beta D_C C_r - \frac{2\alpha}{R(t)}(1-\zeta(t))(A - A^*).$$

Setting

$$\eta(t) = \zeta(t) - 1, \quad q(t) = \frac{2\alpha}{R(t)}(A^* - A(R(t),t)),$$

we can write

$$\dot{\eta} + q(t)\eta = \beta D_C C_r(R(t),t) \equiv p(t), \tag{6.6}$$

so that

$$\eta(t) = e^{-\int_0^t q(\tau)d\tau}\eta(0) + \int_0^t p(\tau)e^{-\int_\tau^t q(s)ds}\,d\tau. \tag{6.7}$$

First consider the case in which there exists a positive constant α_0 and a sequence $t_n \to \infty$ such that

$$\frac{1}{t_n}\int_0^{t_n}(A^* - A(R(\tau),\tau))d\tau \ge \alpha_0 \quad \text{as} \quad t_n \to \infty; \tag{6.8}$$

then,

$$\int_0^{t_n} q(\tau)d\tau \geq \frac{2\alpha}{R_0}\,\alpha_0 t_n \quad \text{as} \quad t_n \to \infty.$$

By using (6.3) in (6.7), we get

$$\begin{aligned}\eta(t_n) &\geq \eta(0)e^{-\frac{2\alpha\alpha_0}{R_0}\,t_n} + \int_{t_0}^{t_n} \frac{M}{\tau^6}\, e^{-\frac{2\alpha A^*}{R_*}(t_n-\tau)}d\tau \\ &\geq \eta(0)e^{-\frac{2\alpha\alpha_0}{R_0}\,t_n} + \frac{1}{2}\,\frac{R_*}{2\alpha A^*}\,\frac{M}{t_n^6}\end{aligned}$$

for some $t_0 > 0$ and all t_n sufficiently large. This is a contradiction since $\eta(t) = \zeta(t) - 1 < 0$ for all $t > 0$.

It remains to consider the case where (6.8) is not satisfied for any $\alpha_0 > 0$, that is,

$$\frac{1}{t}\int_0^t A(R(s),s)ds \to A^* \quad \text{as} \quad t \to \infty. \tag{6.9}$$

Since

$$A_r(R(t),t) = \frac{1}{\alpha}\,\dot{R}(t) \to 0$$

by Lemma 6.1, we can derive, by comparison,

$$A(r,t) \leq \widetilde{W}(r,t) \tag{6.10}$$

for any $\varepsilon > 0$ and $t > t_\varepsilon$, where $\widetilde{W}$ is the supersolution

$$\widetilde{W}(r,t) = \frac{M_\varepsilon}{t^{3/2}}\, e^{-\frac{r^2}{4t\sqrt{D_C}}} + \frac{\varepsilon}{r}(R(t_\varepsilon))^2 \tag{6.11}$$

and t_ε is sufficiently large, so that $\frac{1}{2}\,D_A|\dot{R}(t)| < \varepsilon$ if $t > t_\varepsilon$. It follows that

$$A(R(t),t) \to 0 \quad \text{if} \quad t \to \infty,$$

a contradiction to (6.9). □

7. Asymptotic estimates as $K \to \infty$. We now study the behavior of the solution $(A_K, B_K, C_K, R_K, \zeta_K)$ as $K \to \infty$, assuming that

$$D_B = D_A. \tag{7.1}$$

Recall that

$$0 \leq A_K \leq A^*, \quad 0 \leq B_K \leq B^* \tag{7.2}$$

and

$$\frac{\partial}{\partial r}\,A_K \leq 0, \quad \frac{\partial}{\partial r}\,B_K \geq 0, \quad \dot{R}_K(t) \leq 0. \tag{7.3}$$

LEMMA 7.1. *There exists a positive constant M such that*

$$-M \le \frac{\partial}{\partial r} A_K, \quad \frac{\partial}{\partial r} B_K \le M, \quad -\alpha M \le \dot{R}_K(t) \tag{7.4}$$

for all K.

Proof. Consider the function $u = A_K - B_K$. It satisfies

$$\begin{aligned} &u_t = D_A u \quad \text{if} \quad r > R_K(t),\ t > 0, \\ &-\zeta_{K,0}(t) D_A u_r = (1 - \zeta_{K,0}(t))(A^* - A) \quad \text{if} \quad r = R_K(t),\ t > 0, \\ &u(r,0) = A_0(r), \quad r > R_0, \\ &u(\infty, t) = 0, \quad t > 0, \end{aligned} \tag{7.5}$$

where $\zeta_{K,0} = \min\{\zeta_K, 1\}$. Applying the maximum principle to u_r, we deduce that

$$-M \le u_r \le 0, \tag{7.6}$$

where M is a constant independent of K. But then, upon recalling (7.3) as well,

$$(A_{K,r})^2 = A_{K,r} u_r + A_{K,r} B_{K,r} \le A_{K,r} u_r \le M |A_{K,r}|,$$

so that $|A_{K,r}| \le M$. The proof that $|B_{K,r}| \le M$ is similar. Finally, $\dot{R}_K = \alpha A_{K,r} \le -\alpha M$. □

LEMMA 7.2. *There exist positive constants N_0, N independent of K such that, for any $T > 0$,*

$$\int_0^T \int_{R_K(t)}^{\infty} K A_K B_K r^2 dr dt \le N_0 + NT \tag{7.7}$$

for all K.

Proof. Integrating the equation

$$K A_K B_K = D_A \Delta A_K - \frac{\partial}{\partial t} A_K \tag{7.8}$$

over $R_K(t) < r < \infty$, $0 < t < T$ and using Lemma 7.1, (7.7) readily follows. □

In order to obtain uniform Hölder estimates on A_K, B_K, we consider the function $v = A_K B_K$. It satisfies

$$\begin{aligned} &v_t = D_A \Delta v - 2 A_{K,r} B_{K,r} - K(A_K + B_K) v, \quad r > R_K(t), \\ &-\zeta_{K,0}(t) D_A v_r = (1 - \zeta_{K,0}(t))(A^* - A_K) B_K, \quad r = R_K(t), \\ &v(r,0) = A_0(r) B_0(r), \quad r > R_0, \\ &v(\infty, t) = 0, \quad t > 0. \end{aligned} \tag{7.9}$$

LEMMA 7.3. *For any $T > 0$ and any compact set Ω_T in $\{R_K(t) < r < \infty,\ 0 \le t \le T\}$, whose distance to $r = R_K(t)$ is $\ge c_* > 0$, there exists a constant M depending on T and c_* (but not on K) such that*

$$\int_{\Omega_T} [K A_K B_K (A_K + B_K)]^2 r^2 dr dt \le M. \tag{7.10}$$

Proof. Let $\xi(r,t)$ be a cutoff function such that $\xi = 1$ in Ω_T and $\xi = 0$ outside a $((1/2)c_*)$-neighborhood of Ω_T. Multiplying the differential equation for v by $\xi^2 K v r^2$ and integrating, we obtain

$$\int_{R_K(t)}^{\infty} \xi^2 K \frac{v^2}{2}\, r^2 dr + \int_0^T \int_{R_K(t)}^{\infty} \xi^2 K (v_r)^2 r^2 dr dt + \int_0^T \int_{R_K(t)}^{\infty} \xi^2 K^2 v^2 (A_K + B_K) r^2 dr dt \le I,$$

where

$$I = \int_0^T \int_{R_K(t)}^{\infty} \left[|\xi \xi_t| K v^2 + 2\xi |\xi_r|\ |v_r| K v + 2 |A_{K,r} B_{K,r}| \xi^2 K v \right] r^2 dr dt$$

is bounded independently of K by Lemmas 7.1 and 7.2; this implies the assertion (7.10). □

LEMMA 7.4. *Let Ω_T be as in Lemma 7.3. Then there exists a constant M depending on T and c_* (but not on K) such that*

$$\|A_K\|_{C^{1/4,1/8}(\Omega_T)} \le M, \tag{7.11}$$

$$\|B_K\|_{C^{1/4,1/8}(\Omega_T)} \le M. \tag{7.12}$$

That means that the A_K and B_K are uniformly Hölder continuous (in Ω_T) with exponent $1/4$ in r and exponent $1/8$ in t.

Proof. By Lemmas 7.2 and 7.3,

$$v_t - D_A \Delta v \quad \text{is in} \quad L^2(\Omega_T)$$

uniformly in K. By L^2 estimates it then follows that

$$\|v\|_{W_2^{2,1}(\Omega_T)} \le M$$

with a slightly smaller set Ω_T and a larger constant M, both, however, independent of K. By Sobolev's imbedding [11] we then have

$$\|v\|_{C^{1/2,1/4}(\Omega_T)} \le M \tag{7.13}$$

with yet another constant M.

The function $u = A_K - B_K$ satisfies (7.5) and we can apply L^p estimates to deduce that u also satisfies estimate (7.13). Thus, both $A_K B_K$ and $A_K - B_K$ belong to $C^{1/2,1/4}(\Omega_T)$ uniformly in K. Since

$$A_K + B_K = (4v + u^2)^{1/2},$$

it follows that $A_K + B_K$ is in $C^{1/4,1/8}(\Omega_T)$ uniformly in K, and the same then holds for A_K and B_K. □

THEOREM 7.5. *Let $\widetilde{\Omega}_T$ be any compact domain which is contained in $R_K(t) < r < \infty$, $0 \le t < T$ for all K sufficiently large. Then*

$$\lim_{K\to\infty} A_K(r,t)B_K(r,t) = 0 \tag{7.14}$$

uniformly in $(r,t) \in \widetilde{\Omega}_T$.

Proof. By Lemma 7.2,

$$\iint_{\widetilde{\Omega}_T} A_K B_K r^2 dr dt \le \frac{M}{K} \to 0$$

if $K \to \infty$. Since $A_K B_K$ is uniformly Hölder continuous in $\widetilde{\Omega}_T$ with exponent and coefficient independent of K, (7.14) follows. □

8. Asymptotic limits as $K \to \infty$. The estimates of §7 show that, for any sequence $K_n' \to \infty$, there is a subsequence K_n such that, as $K_n \to \infty$,

$$R_{K_n}(t) \to R(t) \in Lip[0,T] \quad \text{in} \quad C^0[0,T], \tag{8.1}$$

$$A_{K_n}(r,t) \to A(r,t) \in C^{1/4,1/8}(Q_T) \quad \text{in} \quad C^0(Q_T), \tag{8.2}$$

$$B_{K_n}(r,t) \to B(r,t) \in C^{1/4,1/8}(Q_T) \quad \text{in} \quad C^0(Q_T), \tag{8.3}$$

$$K_n A_{K_n} B_{K_n} \to f \quad \text{in the sense of weak convergence of measures} \tag{8.4}$$

for any $0 < T < \infty$, where

$$Q_T = \{(r,t); r > R(t),\ 0 \le t < T\},$$

and f is a measure.

From Theorem 7.5 we have

$$A(r,t)B(r,t) = 0. \tag{8.5}$$

The functions A, B both satisfy the equation

$$w_t - D_A \Delta w + f = 0 \quad \text{in} \quad \mathcal{D}'(Q_T), \tag{8.6}$$

whereas the function $u = A - B$ satisfies

$$u_t - D_A \Delta u = 0 \quad \text{in} \quad Q_T \tag{8.7}$$

since each of the functions $A_{K_n} - B_{K_n}$ satisfies this equation.

By (7.3) it follows that $u_r \le 0$ in Q_T, and then, by the strong maximum principle,

$$u_r < 0 \quad \text{in} \quad Q_T. \tag{8.8}$$

It follows that there exists a curve $r = S(t)$ with $S(t) \in C^{1+\nu}[0,T] \cap C^\infty(0,T)$ such that

$$\begin{aligned} &u(r,t) > 0 \quad \text{if} \quad r < S(t), \\ &u(r,t) < 0 \quad \text{if} \quad r > S(t); \end{aligned} \tag{8.9}$$

here ν is as in (1.10), (1.11). Since $A_0 - B_0$ is positive at $r = R_0$ and negative at $r = \infty$,

$$R_0 < S(0) < \infty. \tag{8.10}$$

Take T such that

$$S(t) > R(t) \quad \text{for all} \quad 0 \le t \le T. \tag{8.11}$$

For $r < S(t)$ we have $u > 0$, or $A > B$. Since $AB = 0$, it follows that $B = 0$. Similarly, $A = 0$ if $r > S(t)$; thus

$$\begin{aligned} A(r,t) &= 0 \quad \text{if} \quad r > S(t), \\ B(r,t) &= 0 \quad \text{if} \quad r < S(t). \end{aligned} \tag{8.12}$$

In any closed domain in $\{r < S(t),\ t < T\}$ we have $B = 0$ and then, by (8.6) with $w = B$, $f = 0$. Similarly, $f = 0$ if $r > S(t)$. It follows that f is a measure supported on $r = S(t)$, $0 < t < T$. In particular,

$$A_t - D_A \Delta A = 0 \quad \text{if} \quad r < S(t), \quad 0 < t < T.$$

Since, in addition, $A(S(t),t) = 0$ and $S(t)$ is smooth, regularity results for the heat equation imply that A is in $C^{1+\nu}$ in $R(t) \le r \le S(t)$, $0 \le t \le T$ and in C^∞ in $R(t) \le r \le S(t)$, $0 < t \le T$.

Next, from (8.6) for $w = A$,

$$-\iint A(-\varphi_t - D_A \Delta \varphi) r^2 dr dt = \iint f \varphi r^2 dr dt \tag{8.13}$$

for any test function φ in Q_T. Using the fact that $A = 0$ if $r > S(t)$ and integrating by parts in (8.13), we find that

$$\iint f \varphi r^2 dr dt = -\int_0^T S^2(t) D_A \, A_r(S(t) - 0, t) \varphi(S(t), t) dt.$$

It follows that

$$f(r,t) = D_A A_r(S(t) - 0, t) \delta(r - S(t)). \tag{8.14}$$

THEOREM 8.1. *For any sequence $K_n' \to 0$ there is a subsequence $K_n \to \infty$ such that the solutions of (P_{K_n}) converge to a solution of (P_∞) uniformly in compact subsets of $\{r > R(t),\ \ 0 \le t \le T\}$.*

Proof. We have already proved most of the theorem. Since $S(t) > R(t)$ for $0 \le t \le T$, it follows that A satisfies the two boundary conditions (2.8), (2.9), where $\zeta = \lim_{K \to \infty} \zeta_K$. Using the $C^{1+\mu}$ regularity of A (which one obtains by the same argument as the one for problem (P_K)) as well as the $C^{1+\mu}$ regularity of $R_K(t)$ and the A_K near $r = R_K(t)$, we can deduce that $C_{K_n} \to C$ uniformly near $r = R(t)$ and C satisfies the boundary condition (2.10). The remaining assertions of Theorem 8.1 have already been proved. □

Denote by T_f the supremum of all T's for which (8.11) holds. We claim that

$$\text{if} \quad T_f < \infty \quad \text{then} \quad S(t) \to R(T_f) \quad \text{if } t \to T_f. \tag{8.15}$$

Indeed, if $\overline{\lim}\ S(t) > R(T_f)$ then (8.11) holds for $T > T_f$ (since u is smooth and $u_r < 0$ in $\{r > R(t),\ t > 0\}$). On the other hand, if the limit $S(T_f - 0)$ does not exist then $u_r(r, T_f)$ will vanish on a nonempty interval, contradicting the inequality $u_r < 0$.

In §9 we shall prove uniqueness for problem (P_∞); this implies that the convergence asserted in Theorem 8.1 is not just for a subsequence $K_n \to \infty$ but for all $K \to \infty$.

In §10 we shall prove that $T_f < \infty$, and this will conclude the proof of all the assertions made in Theorems 2.2 and 2.3.

9. Uniqueness for (P_∞). In this section we prove the following theorem.

THEOREM 9.1. *Assume that $D_A = D_B$. Then there exists at most one solution to problem (P_∞).*

Proof. We begin with some remarks on the regularity of any solution (A, B, C, R, S). Consider the function

$$u = \begin{cases} A & \text{if} \quad R(t) \le r \le S(t), \\ -B & \text{if} \quad r > S(t) \end{cases}$$

for all $t < T$, where $T < T_f$, the final time of the solution. Then

$$u_t = D_A \, \Delta u$$

and $u = A$ in a neighborhood of $\{(R(t), t),\ 0 < t < T\}$. From this and the boundary conditions for A and C at $r = R(t)$ we easily deduce, as in earlier sections, that $R(t)$ and $\zeta(t)$ belong to $C^{1+\mu}[0,T]$ and $C(r,t)$ belongs to $C^{1+\mu}$ in $R(t) \le r < S(t)$, $0 \le t \le T$ for some $0 < \mu < 1$, and

$$-M < u_r < 0 \quad \text{for} \quad r > R(t),\ 0 \le t \le T; \tag{9.1}$$

furthermore, the function $S(t)$ defined by

$$r = S(t) \quad \text{if} \quad u(r,t) = 0$$

is in $C^{1+\mu}[0,T]$ with

$$\dot{S}(t) = -\frac{u_t(S(t),t)}{u_r(S(t),t)}.$$

(Note that $u_t \in C^\mu$ with $\mu = \nu$ up to $t = 0$ since A_0 and B_0 belong to $C^{2+\nu}$.) The function $C(r,t)$ is C^∞ off $r = S(t)$. Near $r = S(t)$, the regularity of $C(r,t)$ is the same as the regularity of the special solution of (2.7):

$$\widetilde{C}(|x|,t) = \int_0^t \int_{|y|=R(\tau)} \frac{e^{-\frac{|x-y|^2}{\sqrt{D_C}\ 4(t-\tau)}}}{(4\pi^2 D_C(t-\tau))^{3/2}} \, u_r(S(\tau),\tau) d\sigma_y d\tau. \tag{9.2}$$

This is a single layer potential and, since $u_r(S(\tau),\tau)$ is in C^μ, this potential is in C^β (across $r = S(t)$) for any $0 < \beta < 1$, and its derivative from each side of $r = S(t)$ is uniformly continuous (the proof is similar to the proof of Lemma 1 in [6, p. 217]). Thus, in particular,

$$|C_r|_{L^\infty} \le M', \quad M' \quad \text{const.} \tag{9.3}$$

Now suppose that $(A_i, B_i, C_i, R_i, S_i)$ are solutions of problem (P_∞) for $i = 1, 2$ and set

$$u_i(r,t) = \begin{cases} A_i(r,t) & \text{if} \quad R_i(t) \le r \le S_i(t), \\ -B_i(r,t) & \text{if} \quad S_i(t) < r < \infty \end{cases}$$

for $0 \le t \le T$, where T is such that

$$S_i(t) - R_i(t) \ge \lambda > 0 \quad \text{for} \quad 0 \le t \le T \quad \text{and} \quad i = 1, 2 \tag{9.4}$$

where $x = r - R_i(t)$ and

$$\widetilde{u}(x,t) = \widehat{u}_1(x,t) - \widehat{u}_2(x,t),$$
$$\widetilde{C}(x,t) = \widehat{C}_1(x,t) - \widehat{C}_2(x,t).$$

Set

$$\Omega_\tau = \{0 < x < \infty,\ 0 < t < \tau\} \quad \text{for any} \quad \tau \le T.$$

The functions $\widetilde{u}, \widetilde{C}$ satisfy (for simplicity we take $D_A = D_C = 1$)

(9.5)

$$\widetilde{u}_t = \widetilde{u}_{xx} + \left(\frac{2}{x+R_1} + \dot{R}_1\right)\widetilde{u}_x + \left[\frac{2}{x+R_1} + \dot{R}_1 - \frac{2}{x+R_2} - \dot{R}_2\right]\widetilde{u}_{2,x} \quad \text{in} \quad \Omega_T,$$
$$-\zeta_{1,0}\widetilde{u}_x + (1-\zeta_{1,0})\widetilde{u} = (\zeta_{1,0} - \zeta_{2,0})[\widetilde{u}_{2,x} + \widetilde{u}_2 - A^*] \quad \text{if} \quad x = 0,\ 0 < t < T,$$
$$\widetilde{u}(x,0) = 0, \quad x > 0,$$
$$\widetilde{u}(\infty,t) = 0, \quad 0 < t < T,$$

and

(9.6)

$$\widetilde{C}_t = \widetilde{C}_{xx} + \left(\frac{2}{x+R_1} + \dot{R}_1\right)\widetilde{C}_x + \left[\frac{2}{x+R_1} + \dot{R}_1 - \frac{2}{x+R_2} - \dot{R}_2\right]\widehat{C}_{2,x}$$
$$+\widehat{u}_{2,x}(S_2 - R_2, t)\delta(x + R_2 - S_2) - \widehat{u}_{1,x}(S_1 - R_1, t)\delta(x + R_1 - S_1) \quad \text{in} \quad \Omega_T,$$

$$-\widetilde{C}_x + \gamma(\widehat{C}_1 + \widehat{C}_2)(\widehat{C}_1^2 + \widehat{C}_2^2)\widetilde{C} = 0 \quad \text{if} \quad x = 0,\ 0 < t < T,$$

$$\widetilde{C}(x,0) = 0, \quad x > 0,$$

$$\widetilde{C}(\infty,t) = 0, \quad 0 < t < T.$$

By the maximum principle,

$$\|\widetilde{u}\|_{L^\infty(\Omega_\tau)} \le \|\zeta_1 - \zeta_2\|_{L^\infty(0,\tau)} + N\tau[\|R_1 - R_2\|_{L^\infty(0,\tau)} + \|\dot{R}_1 - \dot{R}_2\|_{L^\infty(0,\tau)}] \tag{9.7}$$

for any $0 < \tau < T$, where N is a constant independent of τ; in what follows we shall denote any such constant by N.

Differentiating the differential equation in (9.5) in x, we obtain a parabolic equation for $\widetilde{u}_x$. Using the maximum principle, we find that

$$\|\widetilde{u}_x\|_{L^\infty(\Omega_\tau)} \le N\|\zeta_1 - \zeta_2\|_{L^\infty(\Omega_\tau)} + N\tau[\|R_1 - R_2\|_{L^\infty(0,\tau)} + \|\dot{R}_1 - \dot{R}_2\|_{L^\infty(0,\tau)}]. \tag{9.8}$$

Since $R_1(0) = R_2(0)$, we also have

$$\|R_1 - R_2\|_{L^\infty(0,\tau)} \le \tau\|\dot{R}_1 - \dot{R}_2\|_{L^\infty(0,\tau)}. \tag{9.9}$$

From the boundary conditions for $\widetilde{u}_i$ at $x = 0$,

$$|\dot{R}_1 - \dot{R}_2| = \alpha|\widehat{u}_{1,x}(0,t) - \widehat{u}_{2,x}(0,t)|$$
$$= \alpha\left|\frac{1-\zeta_{1,0}}{\zeta_{1,0}}(\widehat{u}_1 - A^*) - \frac{1-\zeta_{2,0}}{\zeta_{2,0}}(\widehat{u}_2 - A^*)\right|,$$

so that

$$\|\dot{R}_1 - \dot{R}_2\| \le N\left[\|\widetilde{u}\|_{L^\infty(\Omega_\tau)} + \|\zeta_1 - \zeta_2\|_{L^\infty(0,\tau)}\right]. \tag{9.10}$$

Substituting (9.9), (9.10) into (9.7), (9.8) and choosing $\tau < 1/(2N)$, we get

$$\|\widetilde{u}\|_{L^\infty(\Omega_\tau)} + \|\widetilde{u}_x\|_{L^\infty(\Omega_\tau)} \le N\|\zeta_1 - \zeta_2\|_{L^\infty(0,\tau)}. \tag{9.11}$$

Next, from the definition (1.14), (1.15) of $\zeta_{i,0}, \zeta_i$ we deduce that

$$\|\zeta_1 - \zeta_2\|_{L^\infty(0,\tau)} \le N\tau\|\widetilde{C}(0,t)\|_{L^\infty(0,\tau)} + N\|R_1 - R_2\|_{L^\infty(0,\tau)}, \tag{9.12}$$

and we need to estimate $\widetilde{C}(0,t)$.

Denote by $G(x,y,t,s)$ the Green function for the problem (see [3])

$$w_t = w_{xx} + \left(\frac{2}{x+R_1} + \dot{R}_1\right)w_x \quad \text{in} \quad \Omega_T,$$
$$-w_x + \gamma(\widehat{C}_1 + \widehat{C}_2)(\widehat{C}_1^2 + \widehat{C}_2^2)w = 0 \quad \text{if} \quad x = 0, \quad 0 < t < T,$$
$$w(\infty,t) = 0, \quad 0 < t < T.$$

Then (using (9.6)) we can represent $\widetilde{C}$ in the form

$$\widetilde{C}(|x|,t) = \int_0^t\int_0^\infty G(x,y,t,s)\left[\frac{2}{y+R_1(s)} + \dot{R}_1(s) - \frac{2}{y+R_2(s)} - \dot{R}_2(s)\right]\widehat{C}_{2,y}(y,s)dyds$$
$$+ \int_0^t G(x,S_2(s) - R_2(s),t,s)\widehat{u}_{2,x}(S_2(s) - R_2(s),s)ds$$
$$- \int_0^t G(x,S_1(s) - R_1(s),t,s)\widehat{u}_{1,x}(S_1(s) - R_1(s),t-s)ds$$
$$\equiv J_1 + J_2 + J_3.$$

We can write

$$|J_2(0,t) + J_3(0,t)| \le \int_0^t |G(0,S_2 - R_2,t,s)||\widehat{u}_{2,x}(S_2 - R_2,s) - \widehat{u}_{1,x}(S_1 - R_1,s)|$$
$$+ \int_0^t |\widehat{u}_{1,x}(S_1 - R_1,s)||G(0,S_2 - R_2,t,s) - G(0,S_1 - R_1,t,s)|,$$

where $S_i = S_i(s)$, $R_i = R_i(s)$. The difference in the first integral is equal to

$$[\widehat{u}_{2,x}(S_2 - R_2,s) - \widehat{u}_{2,x}(S_1 - R_1,s)] + \widetilde{u}_x(S_1 - R_1,s).$$

Therefore, if $0 < t < \tau$,

$$|J_2(0,t) + J_3(0,t)| \le N\left[\|S_1 - S_2\|_{L^\infty(0,\tau)} + \|R_1 - R_2\|_{L^\infty(0,\tau)} + \|\widetilde{u}_x\|_{L^\infty(\Omega_\tau)}\right].$$

Using the regularity result (9.3), we can also immediately estimate $|J_1(0,t)|$ by the L^∞ norm of $|R_1 - R_2| + |\dot{R}_1 - \dot{R}_2|$. Hence

$$|\widetilde{C}(0,t)| \le N[\|\dot{R}_1 - \dot{R}_2\|_{L^\infty(0,\tau)} + \|S_1 - S_2\|_{L^\infty(0,\tau)} + \|\widetilde{u}_x\|_{L^\infty(\Omega_\tau)}]. \tag{9.13}$$

Next we need to estimate $S_1 - S_2$. By the mean value theorem,

$$\widehat{u}_1(S_1 - R_1, t) - \widehat{u}_1(S_2 - R_2, t) = \widehat{u}_{1,x}(y,t) \cdot (S_1 - R_1 - S_2 + R_2),$$

where y is a point between $S_1 - R_1$ and $S_2 - R_2$. Since $|u_{1,x}| \ge \nu_0 > 0$ (ν_0 is a constant which depends on the λ in (9.4)), we get

$$\begin{aligned} |S_1(t) - S_2(t)| &\le |R_1 - R_2| + \frac{1}{\nu_0}|\widehat{u}_1(S_1 - R_1, t) - \widehat{u}_1(S_2 - R_2, t)| \\ &= |R_1 - R_2| + \frac{1}{\nu_0}|\widehat{u}_2(S_2 - R_2, t) - \widehat{u}_1(S_2 - R_2, t)| \\ &\le \|R_1 - R_2\|_{L^\infty(0,\tau)} + \frac{1}{\nu_0}\,\|\widetilde{u}\|_{L^\infty(\Omega_\tau)}. \end{aligned}$$

Substituting this into (9.13) and using the result in (9.12), we find, after also using (9.10) and (9.11), that (9.12) becomes

$$\|\zeta_1 - \zeta_2\|_{L^\infty(0,\tau)} \le N\tau\|\zeta_1 - \zeta_2\|_{L^\infty(0,\tau)}.$$

Hence $\zeta_1(t) = \zeta_2(t)$ if $0 \le t \le \tau$, τ small. This implies (by (9.11) and (9.10)) that $u_1 \equiv u_2$, $R_1 \equiv R_2$, and the two solutions coincide if $0 \le t \le \tau$. Similarly, we proceed step-by-step to complete the proof of uniqueness.

10. $T_f < \infty$. In this section we prove the following theorem.

THEOREM 10.1. *If $D_A = D_B$ and $C_0(r) \not\equiv 0$ then $T_f < \infty$.*

We shall denote by T^*_∞ the smallest t (if it exists) such that $\zeta(t) \ge 1$, and we first prove the following lemma.

LEMMA 10.2. *If $T_f = \infty$ then $T^*_\infty < \infty$.*

Proof. Suppose the assertion is not true; then $\zeta(t) < 1$ for all $t \ge 0$. Since $T_f = \infty$ as well, $A(r,t) = u(r,t)$ for all (r,t) in a neighborhood of $\{(R(t),t),\ 0 \le t < \infty\}$ and, consequently,

$$D_A\zeta u_r + (1-\zeta)(u - A^*) = 0 \quad \text{if} \quad r = R(t),\ \ t > 0, \tag{10.1}$$

$$u_r = \alpha\dot{R} \quad \text{if} \quad r = R(t),\ t > 0, \tag{10.2}$$

where $u = \lim_{K\to\infty}(A_K - B_K)$. Set

$$\sigma = \lim_{t\to\infty} \zeta(t); \quad \text{then} \quad \sigma \le 1.$$

If $\sigma < 1$ then we can use the comparison argument as in the proof of (6.10) to deduce that

$$|u(r,t) - u_\infty(r)| \le \widetilde{W}(r,t) \quad \text{for all} \quad t > 0, \tag{10.3}$$

where $\widetilde{W}$ is defined in (6.11) and $u_\infty(r)$ is the harmonic function in $\rho < r < \infty$, where ($\rho = \lim_{t\to\infty} R(t)$) satisfies

$$-D_A\sigma u_{\infty,r}(\rho) + (1-\sigma)(u_\infty(\rho) - A^*) = 0, \quad u_\infty(\infty) = -B^*. \tag{10.4}$$

Next we argue as in Lemma 6.2 (with $t_0 = 0$ in (6.4)) and deduce that

$$C_{K,r}(R_K(t), t) \geq \frac{M}{t^6} \qquad (M > 0, \quad t \geq 1), \tag{10.5}$$

where M is independent of K. Hence

$$C_r(R(t), t) \geq \frac{M}{t^6} \qquad (M > 0, \quad t \geq 1). \tag{10.6}$$

Using this estimate and (10.1), (10.2), we can proceed as in the proof of Theorem 6.3, deriving (6.7) for $\eta(t) = \zeta(t) - 1$ and concluding that, if

$$\frac{1}{t_n}\int_0^{t_n} (A^* - u(R(\tau),\tau))d\tau \geq \alpha_0 > 0 \tag{10.7}$$

for a sequence $t_n \to \infty$, then $\eta(t_n) > 0$ for t_n large enough, which is a contradiction of the assumption that $\zeta(t) < 1$ for all $t > 0$. Hence

$$\frac{1}{t}\int_0^t u(R(\tau),\tau)d\tau \to A^* \quad \text{as} \quad t \to \infty. \tag{10.8}$$

This, together with (10.3), implies that $u_\infty(\rho) = A^*$. Hence u_∞ takes it maximum at $r = \rho$ and, by the maximum principle, $u_{\infty,r}(\rho) < 0$. However, since $u_\infty(\rho) = A^*$, we must also have $u_{\infty,r}(\rho) = 0$ by (10.4), which is a contradiction. We conclude that $\sigma = 1$, i.e., $\zeta(\infty) = 1$, and then, by (10.1),

$$u_r(R(t), t) \to 0 \quad \text{if} \quad t \to \infty. \tag{10.9}$$

By (10.9) and a comparison argument we then get

$$u(r,t) \leq \widetilde{W}(r,t),$$

where $\widetilde{W}$ is defined in (6.11); therefore

$$u(R(t), t) \to 0 \quad \text{if} \quad t \to \infty. \tag{10.10}$$

Next observe that the argument that led from (10.5) to (10.8) is independent of the question of whether $\sigma < 1$ or $\sigma = 1$. Thus (10.8) still holds and this is a contradiction of (10.10). The proof that T^*_∞ is finite is thereby complete. □

Proof of Theorem 10.1. Suppose $T_f = \infty$. Then, by Lemma 10.2, $T^*_\infty < \infty$. Therefore

$$u_r(\overline{R}, t) = 0 \quad \text{if} \quad t > T^*_\infty \quad (\overline{R} = R(T^*_\infty)).$$

Also, since $u_t = D_A\Delta u$ and $u(\infty, t) = -B^*$, it follows by the comparison argument that

$$|u(r,t) + B^*| \leq M\frac{e^{-\frac{r^2}{\sqrt{D_A}\,4t}}}{t^{3/2}} \quad \text{if} \quad t > T^*_\infty.$$

Consequently, $u(r,t) < 0$ for all $r \geq R(t)$, $t \geq T$ provided T is sufficiently large, and this is a contradiction of the assumption that $T_f = \infty$ (which implies that u remains positive for all $t > 0$). □

REFERENCES

[1] G. ASTASITA, *Mass Transfer with Chemical Reaction*, Elsevier, Amsterdam, 1967.

[2] J.R. CANNON AND C.D. HILL, *On the movement of a chemical reaction interface*, Indiana Univ. Math. J., 20 (1970), pp. 429–454.

[3] S.D. EIDELMAN, *Parabolic Systems*, North–Holland, Groningen, 1969.

[4] P. ERDI AND J. TOTH, *Mathematical Models of Chemical Reactions*, Princeton University Press, Princeton, NJ, 1989.

[5] L.C. EVANS, *A convergence theorem for a chemical diffusion-reaction system*, Houston J. Math., 6 (1980), pp. 259–267.

[6] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[7] ———, *One phase moving boundary problems*, in Conference on Moving Boundary Problems, Academic Press, New York, 1978, pp. 25–40.

[8] ———, *Variational Principles and Free-Boundary Problems*, Wiley-Interscience, New York, 1982.

[9] ———, *Mathematics in Industrial Problems, Part* 6, IMA Vol. Math. Appl., Volume 57, Springer-Verlag, New York, 1993.

[10] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1986.

[11] O. A. LADYZHENSKAJA, V. A. SOLONNIKOV, AND N. N. URALTSEVA, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc. Trans. Math. Monographs, Vol. 23, American Mathematical Society, Providence, RI, 1968.

[12] R.H. MARTIN JR., *Mathematical models in gas-liquid reactions*, Nonlinear Anal., 4 (1980), pp. 509–527.

[13] A.E. NIELSEN, *Kinetics of Precipitation*, Pergamon Press, Oxford, 1964.

[14] A. VISINTIN, *Stefan problem with a kinetic condition at the free boundary*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 97–122.

[15] W. XIE, *The Stefan problem with a kinetic condition at the free boundary*, SIAM J. Math. Anal., 21 (1990), pp. 362–373.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1113–1129, September 1995

002

ON INTERFACE CONDITIONS FOR A THIN FILM FLOW PAST A POROUS MEDIUM*

GUY BAYADA† AND MICHÈLE CHAMBAT‡

Abstract. This paper is concerned with the problem of finding the boundary conditions at the interface between a fluid flow in a porous medium and an adjacent narrow layer of free fluid flow in, for instance, a lubrication area or cracks in a porous medium. It is a two-small-parameter problem, which can be rigorously studied using both asymptotic analysis and homogenization theory. The existence of a critical ratio between the characteristic length of the porous medium microtexture and the characteristic width of the gap is proved. Interface conditions are found to be the continuity of the pressure and of the normal velocity while the tangential velocity vanishes. The existence and uniqueness of the coupled limit problem is obtained.

Key words. Stokes flow, thin film Reynolds equation, homogenization, asymptotic analysis, interface conditions

AMS subject classifications. 41A60, 76D07, 76S05, 76D08

1. Introduction.

This paper is concerned with the study of a fluid flow in a porous medium drawn by an adjacent narrow layer of free fluid flow whose motion is imposed through nonhomogeneous boundary conditions. The precise situation described here is related to a lubricated device such as a porous journal bearing.

Under the periodicity assumption of the microstructure of the porous medium, it is known [21] that the macroscopic behavior of the flow obeys the Darcy law. In the thin film flow, the asymptotic pressure follows the Reynolds thin film equation which is related to the height of the thin layer and to the velocity on the boundary.

The aim of the present study is to find the boundary condition that should be imposed at the interface between the two media. Different approaches exist. If the thin film aspect is neglected, we can refer to [2], [17], and [20] in which the continuity of the normal flux and a nonslip condition for the tangential velocity are proposed. The problem has been restudied in [8] and [12], which take into account multiple scales and asymptotic expansions. If the thin film thickness is taken into account, the situation is still ambiguous. At first glance, the thin thickness of the free fluid could emphasize the roughness effects of the porous medium, but to what extent? Moreover, experimental evidence is overshadowed by the occurrence of other phenomena such as cavitation [3], [13]. Most authors have used a continuity boundary condition in which the tangential velocity at the interface for the free fluid flow is chosen to be the same as the tangential one issued from the porous medium using the Darcy law [13], [16], [19]. See also [9] and [18].

The main result of the present work is that there is a continuity of the pressure at the interface while the tangential velocity vanishes. This result is obtained for a critical ratio between the height of the free fluid film and the depth of the microstructure of the porous medium. For this critical ratio, the two asymptotic pressures have the same

* Received by the editors April 2, 1993; accepted for publication (in revised form) February 15, 1994.

†Département de Mathématiques, Institut National des Sciences Appliqués de Lyon, Batiment 401, 69621 Villeurbanne cedex, France.

‡Laboratoire d'Analyse Numerique, Université de Lyon I, Batiment 101, 69622 Villeurbanne cedex, France.

order of magnitude at each side of the interface, leading to a balance of the fluxes issuing from each of the media. Similar critical ratios have already been exhibited in other situations, like thin cracks through porous media in fluid problems [4], [11] or in thermal ones [5].

Section 2 is devoted to the introduction of the two-small-parameter problem. In §3, both limit problems at each side of the interface are obtained separately. Coupling effects are introduced in §4, allowing the determination of the critical ratio and the general features of the coupled limit problem. To obtain a well-defined asymptotic equation, further results are needed which relate to the behavior of the averages of the pressure in each of the domains. This is done in §5. Finally, by introducing an intermediary layer inside the free film flow, we are able to prove in §6 the continuity of the asymptotic pressure at the interface. As a consequence, we show the existence and uniqueness of the limit problem, which appears to be a Ventcel-type problem [10] in which a Reynolds equation acts as a boundary condition for the Darcy problem.

2. The basic problem. Ω is a bounded domain in R^3—a cube, for instance. At a microscopic level, $\Omega = \Omega^\varepsilon \cup T^\varepsilon$, where the fluid part Ω^ε is obtained by subtracting from Ω a set T^ε of similar obstacles of size $\varepsilon, \varepsilon \ll 1$, periodically deduced by translation and homothetic transformation from a given obstacle T. We assume that T is a connected set with a smooth boundary strictly contained in the unitary cube $Y = [0,1]^3$. We denote by $\Sigma = \omega \times \{0\}$ the upper surface of Ω and let $\Gamma = \partial\Omega \setminus \Sigma$.

The thin film I^η where the free fluid lies is bounded below by Σ and bounded above by an upper surface Γ_+^η and a lateral surface Γ_s^η. We obtain

$$X \in I^\eta \Leftrightarrow x \in \omega, \quad 0 \le x_3 \le \eta h(x),$$
$$X \in \Gamma_s^\eta \Leftrightarrow x \in \partial\omega, \quad 0 \le x_3 \le \eta h(x),$$

with $X = (x_1, x_2, x_3)$ and $x = (x_1, x_2)$.

The smooth function $h(x)$ describes the shape of the fluid film and η is a small parameter. We assume that

$$0 < a \le h(x) \le b \quad \forall x \in \omega.$$

In the domain $D^{\varepsilon\eta} = \Omega^\varepsilon \cup I^\eta \cup \Sigma$, the fluid obeys the Stokes law where inertial terms have been neglected, since the Reynolds number is small [13]:

$$\mu \Delta u^{\varepsilon\eta} = \nabla p^{\varepsilon\eta}, \tag{2.1}$$
$$\operatorname{div} u^{\varepsilon\eta} = 0 \quad \text{in } D^{\varepsilon\eta}, \tag{2.2}$$

where $u^{\varepsilon\eta}$ and $p^{\varepsilon\eta}$ are the velocity and pressure, while μ is the viscosity.

Boundary conditions are given on the velocity such that there exists a smooth lift $g^{\varepsilon\eta}$ in $H^{1/2}(\partial D^{\varepsilon\eta})^3$ so that

$$g^{\varepsilon\eta} = 0 \quad \text{on } \partial T^\varepsilon \cup \Gamma, \tag{2.3}$$
$$g^{\varepsilon\eta} = s^\eta = (s_1, s_2, \eta s_3) \quad \text{on } \Gamma_+^\eta, s_1, s_2, s_3 \text{ given in } R, \tag{2.4}$$
$$g^{\varepsilon\eta} = (g_1^\eta, g_2^\eta, \eta g_3^\eta) \quad \text{on } \Gamma_s^\eta, g^\eta \text{ a given function.} \tag{2.5}$$

Condition (2.3) is a nonslip condition on the solid boundaries, while (2.4) is related to a given velocity field for the upper part of the device.

Taking into account the small height of the free flow, we introduce the following rescaling in the thin film:

$$z = x_3/\eta \quad \text{for } x_3 > 0, \tag{2.6}$$

so that all geometrical data with superscript η are associated with fixed geometrical elements (with respect to η), which will be marked by omitting the superscript η; for instance $I, D^\varepsilon, \ldots$ corresponds to $I^\eta, D^{\varepsilon\eta}$. We assume that

$$g^\eta(x, \eta z) = g(x, z), \quad 0 < z < h(x), \quad x \in \partial\omega, \tag{2.7}$$

where $g(x, z)$ is a given function independent of η, and we suppose a compatibility condition for the boundary condition

$$\int_{\Gamma_s} g \cdot n \, d\sigma - \int_\omega s \cdot \nabla h \, dx + \int_\omega s_3 \, dx = 0, \tag{2.8}$$

where $s = (s_1, s_2)$ and $n = (n_1, n_2, n_3)$ is the unit normal vector exterior to I, so that (2.1)–(2.5) is replaced by

$$\begin{aligned} &u^{\varepsilon\eta} \in V^{\varepsilon\eta}, \qquad p^{\varepsilon\eta} \in L^2(D^{\varepsilon\eta}), \\ &\mu \int_{D^{\varepsilon\eta}} \nabla u^{\varepsilon\eta} \cdot \nabla \Phi \, dX = \int_{D^{\varepsilon\eta}} p^{\varepsilon\eta} \operatorname{div} \Phi \, dX \quad \forall \Phi \in H_0^1(D^{\varepsilon\eta})^3, \qquad (2.9) \\ &\int_{D^{\varepsilon\eta}} q \operatorname{div} u^{\varepsilon\eta} \, dX = 0 \quad \forall q \in L^2(D^{\varepsilon\eta}), \qquad (2.10) \\ &V^{\varepsilon\eta} = \{\Phi \in H^1(D^{\varepsilon\eta})^3, \Phi = g^\eta \text{ on } \partial D^{\varepsilon\eta}\}, \end{aligned}$$

whose solution is unique [22], the pressure being determined up to an additive constant.

3. Asymptotic analysis. The aim of this section is to obtain the asymptotic behavior of the pressure and the velocity in each of the subdomains. An interesting feature of this study is that no assumptions are made about the relative behavior of ε and η. In §3.2, we find the classical Reynolds limit equation in the thin film and in §3.3 the Darcy law in the porous medium. Coupling effects only appear because of the existence of unknown terms in both limit equations.

In the following, $|\ |_A$ will denote the $L^2(A)$ or the $L^2(A)^3$ norms, while K will be any constant with respect to η and ε. Rescaling (2.6) induces a fixed domain I and we will denote $dI = dxdz$.

For any function ϕ in $L^2(I)$, the average through the gap is written as

$$\bar{\phi}(x) = \frac{1}{h(x)} \int_0^{h(x)} \phi(x, z) \, dz.$$

3.1. A priori estimates. Assumptions (2.7) and (2.8) allow us to find an extension of the nonhomogeneous boundary conditions, whose divergence is zero and whose behavior with respect to ε and η is known.

$$\begin{aligned} &\exists J \in H^1(I)^3, \quad \operatorname{div}(J) = 0 \quad \text{in } I, \\ &\qquad\qquad J = g \quad \text{on } \Gamma_+ \cup \Gamma_s, \\ &\qquad\qquad J = (0, 0, 0) \quad \text{on } \Sigma. \end{aligned} \tag{3.1}$$

LEMMA 3.1. *The following estimates are valid:*

$$\left|\frac{\partial u^{\varepsilon\eta}}{\partial z}\right|_I \le K, \qquad \eta\left|\frac{\partial u^{\varepsilon\eta}}{\partial x_i}\right|_I \le K, \quad i=1,2, \tag{3.2}$$

$$\sqrt{\eta}|\nabla u^{\varepsilon\eta}|_\Omega \le K. \tag{3.3}$$

Proof. Setting $G^\eta = (J_1, J_2, \eta J_3)$ in I^η and zero on Ω, where J is given by (3.1), $u^{\varepsilon\eta} - G^\eta$ is a test function for (2.9). Rescaling (2.6) and extending the velocity by zero in the whole Ω, we get

$$\begin{aligned}\int_\Omega (\nabla u^{\varepsilon\eta})^2\,dX &+ \int_I \eta \sum_{i=1}^{3}\sum_{j=1}^{2}\left(\frac{\partial u_i^{\varepsilon\eta}}{\partial x_j}\right)^2 dI + \int_I \frac{1}{\eta}\sum_{i=1}^{3}\left(\frac{\partial u_i^{\varepsilon\eta}}{\partial z}\right)^2 dI \\ &= \int_I \eta \sum_{i,j=1}^{2}\frac{\partial u_i^{\varepsilon\eta}}{\partial x_j}\frac{\partial J_i}{\partial x_j}dI + \int_I \frac{1}{\eta}\sum_{i=1}^{2}\frac{\partial u_i^{\varepsilon\eta}}{\partial z}\frac{\partial J_i}{\partial z}dI + \int_I \eta^2 \sum_{j=1}^{2}\frac{\partial u_3^{\varepsilon\eta}}{\partial x_j}\frac{\partial J_3}{\partial x_j}dI \\ &\quad + \int_I \frac{\partial u_3^{\varepsilon\eta}}{\partial z}\frac{\partial J_3}{\partial z}dI.\end{aligned}$$

As J does not rely on ε, using the Cauchy–Schwarz inequality, the estimates of the lemma are proved. □

3.2. Limit equation in the thin film I. The asymptotic analysis of the Stokes thin film is performed using the same tools as those for an impervious boundary [1].

PROPOSITION 3.2. *There exists u^* in Vz, s^* in $L^2(\omega)^3$, and a subsequence of $(u^{\varepsilon\eta})$ again denoted by $(u^{\varepsilon\eta})$, such that as ε and η tend to $0, u^{\varepsilon\eta}$ weakly converges towards u^* in Vz such that*

$$u_3^* = 0, \tag{3.4}$$

$$u^* = s \quad \text{on } \Gamma_+, \tag{3.5}$$

$$u_{i/\Sigma}^{\varepsilon\eta} \text{ tends to } s_i^* \text{ weakly in } L^2(\omega) \quad \text{for } i=1,2, \tag{3.6}$$

where $Vz = \{\phi \in L^2(I)^3$ such that $\partial\phi/\partial z \in L^2(I)^3\}$.

Proof. $u^{\varepsilon\eta}$ has nonhomogeneous boundary conditions on Γ_+, but these conditions are independent of ε and η. So, after a translation, we use Friedrich's inequality in the z direction and estimates of Lemma 3.1 show that $u^{\varepsilon\eta}$ is bounded in Vz. The classical compactness argument allows us to extract a subsequence of $u^{\varepsilon\eta}$ which weakly converges in Vz. By using the continuity of the trace from Vz to $L^2(\Sigma)^3$ and $L^2(\Gamma_+)^3$ (see [7]), (3.5) and (3.6) are obtained. Moreover, (2.4) shows that $u_3^* = 0$ on Γ_+. As $u^{\varepsilon\eta}$ converges in $D'(I)$ and using (2.2), we obtain

$$\frac{\partial u_3^*}{\partial z} = \lim \eta\left(\frac{\partial u_1^{\varepsilon\eta}}{\partial x_1} + \frac{\partial u_2^{\varepsilon\eta}}{\partial x_2}\right) = 0.$$

Then the zero value of u_3^* at the boundary implies that $u_3^* = 0$ in the whole I. □

PROPOSITION 3.3. *There exists a subsequence of $u^{\varepsilon\eta}$ such that the trace of $u^{\varepsilon\eta}/\eta$ on Σ weakly converges towards t_3^* in $H^1(\omega)'$ satisfying the following equality. For any ϕ in $H^1(\omega)$,*

$$\langle t_3^*, \phi\rangle = -\sum_{i=1}^{2}\int_\omega \left(\frac{\partial\phi}{\partial x_i}h\bar{u}_i^* + s_i\frac{\partial h}{\partial x_i}\phi\right)dx + \int_\omega \phi s_3\,dx + \int_{\partial\omega} h\bar{g}\cdot n\phi\,d\gamma; \tag{3.7}$$

$\langle , \rangle$ *denotes the duality between* $H^1(\omega)$ *and* $H^1(\omega)'$.

Proof. Let $\phi(x)$ be any function in $H^1(\omega)$. We define a test function q in (2.10) such that $q(x,z) = \phi(x)$ in I and $q = 0$ in Ω^ε, and we obtain

$$\int_{I^\eta} \operatorname{div}(u^{\varepsilon\eta})\phi \, dX = 0.$$

Performing the rescaling (2.6) and using the Green formula, this equality reduces to:

$$\sum_{i=1}^{2} \left(-\int_I \frac{\partial \phi}{\partial x_i} u_i^{\varepsilon\eta} \, dI + \int_{\partial I} \phi u_i^{\varepsilon\eta} n_i \, d\sigma \right) + \frac{1}{\eta} \int_{\partial I} \phi u_3^{\varepsilon\eta} n_3 \, d\sigma = \frac{1}{\eta} \int_I \frac{\partial \phi}{\partial z} u_3^{\varepsilon\eta} \, dI = 0.$$

All values of the velocities are known on ∂I except on Σ, so that the equality is rewritten

$$\begin{aligned} &-\sum_{i=1}^{2} \left(\int_I \frac{\partial \phi}{\partial x_i} u_i^{\varepsilon\eta} \, dx \, dz - \int_\omega \phi(x) s_i \frac{\partial h}{\partial x_i} \, dx \right) \\ &+ \sum_{i=1}^{2} \int_{\Gamma_s} \phi g_i n_i \, d\sigma + \frac{1}{\eta} \int_\omega \phi \eta s_3 \, dx = \frac{1}{\eta} \int_\omega \phi u_3^{\varepsilon\eta} \, dx. \end{aligned} \tag{3.8}$$

Now using Lemma 3.1 and Proposition 3.2, we get

$$\left\| \frac{u_3^{\varepsilon\eta}}{\eta} \right\|_{H^1(\omega)'} = \frac{\operatorname{Sup} \left| \frac{1}{\eta} \int_\omega u_3^{\varepsilon\eta} \phi \, dx \right|}{|\phi|_{H^1(\omega)}} \leq K \frac{(|\nabla \phi|_I + |\phi|_I)}{\|\phi\|_{H^1(\omega)}}.$$

As ϕ does not rely upon z, we obtain

$$\left\| \frac{u_3^{\varepsilon\eta}}{\eta} \right\|_{H^1(\omega)'} \leq K.$$

Taking the limit in (3.8) by using Proposition 3.2 and averaging through the gap, we get (3.7). □

PROPOSITION 3.4. *Let* $p_+^{\varepsilon\eta}$ *be the solution of* (2.9), (2.10) *with* $\int_I p_+^{\varepsilon\eta} \, dI = 0$; *there exists a subsequence* $p_+^{\varepsilon\eta}$ *such that* $\eta^2 p_+^{\varepsilon\eta}$ *weakly converges in* $L^2(I)$ *towards* p_+^* *with* $\partial p_+^*/\partial z = 0$.

Proof. Let ϕ be any function in $H_0^1(I)$. We choose $\Phi = (\phi, 0, 0)$ in (2.9), then $(0, \phi, 0)$ and $(0, 0, \phi)$. The estimates in Lemma 3.1 give

$$\eta^2 \left\| \frac{\partial p_+^{\varepsilon\eta}}{\partial x_i} \right\|_{H^{-1}(I)} \leq K, \qquad i = 1, 2, \tag{3.9}$$

$$\eta \left\| \frac{\partial p_+^{\varepsilon\eta}}{\partial z} \right\|_{H^{-1}(I)} \leq K. \tag{3.10}$$

Due to the particular choice of the pressure determination $p_+^{\varepsilon\eta}$ which lies in the zero-average closed subspace of $L^2(I)$, (3.9) and (3.10) imply that (see [22]) $|\eta^2 p_+^{\varepsilon\eta}|_I$ is bounded, which gives the convergence of $\eta^2 p_+^{\varepsilon\eta}$ to p_+^* weakly in $L^2(I)$, while the comparison between (3.9) and (3.10) directly implies that $\partial p_+^*/\partial z = 0$. □

The goal is now to express u_i^* in terms of p_+^* in the divergence-free equation to obtain the characteristic Reynolds equation for the pressure in the thin film.

THEOREM 3.5. *The pressure obeys the Reynolds equation in* $H^1(\omega)$:

$$\sum_{i=1}^{2}\int_\omega \left[\frac{h^3}{12\mu}\frac{\partial p_+^*}{\partial x_i} - (s_i + s_i^*)\frac{h}{2}\right]\frac{\partial \phi}{\partial x_i}\,dx = \sum_{i=1}^{2}\int_\omega \left(\phi s_i \frac{\partial h}{\partial x_i} - \phi s_3\right)dx$$

$$-\sum_{i=1}^{2}\int_{\partial\omega} h\phi\bar{g}_i\cdot n_i\,d\gamma + \langle t_3^*, \phi\rangle \quad \forall \phi \in H^1(\omega), \quad p_+^* \in H^1(\omega),$$

with the supplementary condition $\int_\omega h p_+^*\,dx = 0$.

Proof. Let $\theta(x)$ be any function in $H_0^1(\omega)$. Choosing $\phi = (z(z-h)\theta(x), 0, 0)$ in I and 0 in Ω (ϕ is a test function in (2.9)) and taking the limit, we get

$$I_1 = \int_I \mu\frac{\partial u_1^*}{\partial z}(2\theta z - \theta h)\,dI = \int_I p_+^*\left(z(z-h)\frac{\partial\theta}{\partial x_1} - \theta z\frac{\partial h}{\partial x_1}\right)dI = I_2.$$

Applying Green's formula in the z-direction as u_1^* lies in Vz and using the boundary values, the first integral becomes

$$I_1 = -2\int_I \mu u_1^*\theta\,dI + \int_{\partial I}\mu u_1^*(2\theta z - \theta h)n_3\,d\sigma = \int_\omega \mu(-2\bar{u}_1^* + s_1^* + s_1)h\theta\,dx.$$

Now applying the Fubini theorem to the right-hand side I_2 where p_+^* does not rely on z, we obtain

$$I_2 = -\int_\omega p_+^*\left(\frac{h^3}{6}\frac{\partial\theta}{\partial x_1} + \frac{h^2}{2}\frac{\partial h}{\partial x_1}\theta\right)dx = \left\langle \frac{\partial p_+^*}{\partial x_1}, \frac{h^3}{6}\theta\right\rangle.$$

From the equality $I_1 = I_2$, we obtain

$$\frac{h^3}{6\mu}\frac{\partial p_+^*}{\partial x_1} = (-2\bar{u}_1^* + s_1^* + s_1)h \quad \text{in } H^{-1}(\omega). \tag{3.11}$$

A similar result is obtained for $\partial p_+^*/\partial x_2$ by choosing $(0, z(z-h)\theta, 0)$ as a test function in (2.9).

As the right-hand side is not only in $H^{-1}(\omega)$ but also in $L^2(\omega)$, (3.11) gives a supplementary regularity result for p_+^*, which in turn is in $H^1(\omega)$ and not only in $L^2(\omega)$. Substituting $\bar{u}_i^*$ for (3.11) in (3.7) and performing Green's integration enables the result to be obtained. □

3.3. Limit problem in the porous media. As in the classic homogenization of the Stokes flow, it is necessary to obtain estimates of the velocity and pressure in the whole domain Ω. Because of (2.3), we extend $u^{\varepsilon\eta}$ by zero to the whole Ω and we use the same notation. We will use the extension of Tartar [21] for the pressure, which is determined regardless of the constant that is added.

LEMMA 3.6. *There exists a linear operator* $R^\varepsilon \in L(H^1(\Omega)^3, H^1(\Omega^\varepsilon)^3)$ *so that*

$$R^\varepsilon(w) = w \quad \text{on } \partial\Omega,$$

$$R^\varepsilon(w) = 0 \quad \text{on } \partial T^\varepsilon,$$

$$w = 0 \quad \text{on } \partial T^\varepsilon \Rightarrow R^\varepsilon(w) = w,$$

$$\operatorname{div}(w) = 0 \Rightarrow \operatorname{div}(R^\varepsilon(w)) = 0,$$

$$|R^\varepsilon(w)|_{\Omega\varepsilon} \le K(|w|_\Omega + \varepsilon|\nabla w|_\Omega), \tag{3.12}$$

$$|\nabla R^\varepsilon(w)|_{\Omega\varepsilon} \le K\left(\tfrac{1}{\varepsilon}|w|_\Omega + |\nabla w|_\Omega\right). \tag{3.13}$$

Proof. The proof is that of Lemma 4 in [21, p. 373].

It is now possible to build an extension for the pressure.

LEMMA 3.7. *There exists an extension of the pressure* $p^{\varepsilon\eta}$, *denoted by* $p_-^{\varepsilon\eta}$ *with zero average on* Ω *so that*

$$|p_-^{\varepsilon\eta}|_\Omega \leq \frac{K}{\varepsilon\sqrt{\eta}}. \tag{3.14}$$

Proof. Let us define F^ε in $H^{-1}(\Omega)$ by

$$\langle F^\varepsilon, w\rangle = \langle \nabla p^{\varepsilon\eta}, R^\varepsilon(w)\rangle \quad \forall w \in H_0^1(\Omega),$$

where $\langle , \rangle$ is the duality product between $H^{-1}(\Omega)^3$ and $H_0^1(\Omega)^3$, and R^ε is defined in Lemma 3.6. Using (2.9), we obtain

$$\langle F^\varepsilon, w\rangle = -\int_{\Omega^\varepsilon} \mu \nabla u^{\varepsilon\eta} \nabla R^\varepsilon(w)\, dX,$$

so that $\langle F^\varepsilon, w\rangle$ is zero if $\operatorname{div}(w) = 0$ and F^ε appears to be the gradient of a function $q^{\varepsilon\eta}$ in $L^2(\Omega)$ such that $\langle \nabla q^{\varepsilon\eta}, w\rangle = \langle \nabla p^{\varepsilon\eta}, w\rangle$ as soon as w is in $H_0^1(\Omega^\varepsilon)$. Then $q^{\varepsilon\eta}$ is an extension of $p^{\varepsilon\eta}$, which is defined as whatever additive constant is added. In the following, we will denote by $p_-^{\varepsilon\eta}$ the particular choice of $q^{\varepsilon\eta}$ with a zero-average value on Ω :

$$\int_\Omega p_-^{\varepsilon\eta}\, dX = 0. \tag{3.15}$$

From (3.3) and (3.13), we get

$$|\nabla p_-^{\varepsilon\eta}|_{H^{-1}(\Omega)} = |\nabla q^{\varepsilon\eta}|_{H^{-1}(\Omega)} \leq K/\varepsilon\sqrt{\eta},$$

and (3.14) directly follows as $p_-^{\varepsilon\eta}$ is zero-average. □

PROPOSITION 3.8. *There exist subsequences such that when* ε *and* η *tend to zero,*
$\varepsilon\sqrt{\eta}p_-^{\varepsilon\eta}$ *strongly tends towards a function* p_-^* *in* $L^2(\Omega)$,
$(\sqrt{\eta}/\varepsilon)u^{\varepsilon\eta}$ *weakly tends towards a function* v^* *in* $H(\operatorname{div}, \Omega)$,
v^* *and* p_-^* *fulfill the Darcy law*

$$v^* = -\mathbf{K}\nabla p_-^* \quad \text{in } H^{-1}(\Omega), \tag{3.16}$$

$$\operatorname{div}(v^*) = 0, \tag{3.17}$$

$$\langle v^* \cdot n, \phi\rangle = 0 \quad \forall \phi \in H^{1/2}(\partial\Omega), \quad \phi = 0 \text{ on } \Sigma. \tag{3.18}$$

$\mathbf{K}$ *is the permeability matrix defined by* (3.21).

Proof. Classical homogenization techniques [21] using the Poincaré inequality in Ω^ε give the following estimates:

$$|u^{\varepsilon\eta}| \leq K\varepsilon|\nabla u^{\varepsilon\eta}|_\Omega \leq K\varepsilon/\sqrt{\eta}. \tag{3.19}$$

But as $\operatorname{div} u^{\varepsilon\eta} = 0$, then $(\sqrt{\eta}/\varepsilon)u^{\varepsilon\eta}$ is bounded in $H(\operatorname{div}, \Omega)$ and (3.17) is obtained. The continuity of the trace from $H(\operatorname{div}, \Omega)$ to $H^{-1/2}(\partial\Omega)$ and the homogeneous boundary condition for $u^{\varepsilon\eta}$ on $\partial\Omega \setminus \Sigma$ imply (3.18).

To get (3.16), the energy method works exactly as in [21]. As it will be useful for §5, we restate some of the proof here.

Letting (e^i) be the canonical basis in R^3, we define the so-called local problem in the following y-variables. Find w^i, Π^i, the y-periodic solutions of the nonhomogeneous Stokes problem

$$\begin{aligned} \mu\Delta_Y w^i &= \nabla_Y \Pi^i - e^i \quad \text{in } Y \setminus T, \\ \operatorname{div}_Y w^i &= 0 \quad \text{in } Y \setminus T, \quad i = 1,2,3, \\ w^i &= 0 \quad \text{on } \partial T. \end{aligned} \tag{3.20}$$

These problems have a unique solution in $H^1(Y \setminus T)^3$ and $L^2(Y \setminus T)$ and are used to define the components of the permeability matrix $\mathbf{K}$:

$$\mathbf{K}_{ij} = \int_{Y\setminus T} (w^i)_j \, dy, \qquad j = 1,2,3. \tag{3.21}$$

Due to the regularity of T, we have

$$w^i \in H^2(Y \setminus T)^3 \quad \text{and} \quad \Pi^i \in H^1(Y \setminus T). \tag{3.22}$$

Now, for any ϕ in $H_0^1(\Omega)$, ϕw^i is a test function for (2.9) while $\phi u^{\varepsilon\eta}$ is one for the weak formulation associated with (3.20). Comparing both expressions and letting ε and η tend to zero leads to (3.16). □

As v^* is in $L^2(\Omega)^3$, we deduce the following theorem from (3.15)–(3.17).

THEOREM 3.9. *p_-^* is in $H^1(\Omega)$ and satisfies the elliptic equation*

$$\begin{aligned} &\operatorname{div}\mathbf{K}\nabla p_-^* = 0 \quad \text{in } H^{-1}(\Omega), \\ &\mathbf{K}\nabla p_-^* \cdot n = v^* \cdot n \quad \text{in } H^{-1/2}(\partial\Omega). \end{aligned}$$

4. Coupling effect. Until now, each domain has been considered separately and the limit problems given in Theorems 3.5 and 3.9 are based on two unknowns, t_3^* and v_3^*, on the interface. In this section, we show that a critical ratio $\varepsilon = O(\eta^{3/2})$ involves an equilibrium between the two flows at Σ.

4.1. Calculation of s^*. We will prove that for small ε, the influence of the porous medium induces a zero limit velocity on the interface. This is a consequence of the following lemma.

LEMMA 4.1. *The following estimate is valid:* $\sqrt{\eta/\varepsilon}|u^{\varepsilon\eta}|_\omega \le K$.

Proof. Following Nguetseng [15], we may use the Poincaré norm for $u^{\varepsilon\eta}$ in the ε-layer of basic cells in the porous medium defined by

$$Q^\varepsilon = \{X, x \in \omega, -\varepsilon \le x_3 \le 0\}.$$

In the basic cell, we obtain in the y-variables

$$\int_{\Sigma\cap\partial Y} (u^{\varepsilon\eta})^2 \, dy_1 \, dy_2 \le K\|u^{\varepsilon\eta}\|^2_{H^{1/2}(\partial Y)^2} \le K|\nabla_y u^{\varepsilon\eta}|^2_Y.$$

After rescaling in the x macrovariables and summation in Q^ε, this inequality becomes

$$|u^{\varepsilon\eta}|^2_\omega \le K\varepsilon \int_{Q^\varepsilon} (\nabla u^{\varepsilon\eta})^2 \, dX \le K\varepsilon|\nabla u^{\varepsilon\eta}|^2_\Omega.$$

The proof ends by using estimates (3.3). □

From (3.6), we immediately obtain the following corollary.

COROLLARY 4.2. *For* $\varepsilon = o(\eta), s^* = 0$.

4.2. The critical case. We recall now that convergences occur in each domain concerning different powers with respect to ε and η, especially those of the normal velocity components which are obtained by the Vz and $H(\mathrm{div}, \Omega)$ convergences of the velocity in I and Ω, respectively.

The physical aspect of the coupling is to impose conditions so that in the two media, the convergences occur simultaneously with the same order of magnitude, inducing a balance of the two normal fluxes through the interface. This will only take place for a particular ratio between the small parameters: we call it the critical ratio and cancel the superscript η in the notation when in that situation.

To give a mathematical meaning to the determination of such a ratio, we first need test functions with additional regularity on Σ. This kind of function has already been used for solving thin-coating elasticity problems in the Ventcel problem [10] and for the study of some nonlinear free boundary problems related to porous bearings [3], [6].

Let $H^1_\Sigma = \{\phi, \phi \in H^1(\Omega), \phi/_\Sigma \in H^1(\omega)\}$.

LEMMA 4.3. H^1_Σ *is a Hilbert space with respect to the norm* $\| \quad \|_\Sigma$:

$$\|\phi\|^2_\Sigma = \|\phi\|^2_{H^1(\Omega)} + \|\phi\|^2_{H^1(\omega)}.$$

Proof. As this space may be equivalently defined by the closure of $D(\Omega)$ with respect to the norm $\| \quad \|_\Sigma$, the result is obvious. □

PROPOSITION 4.4. *For* $\varepsilon = O(\eta^{3/2})$, *the limit problem for* (2.9) *and* (2.10) *is defined in this way (for the sake of simplicity, we take* $O(\eta^{3/2}) = \eta^{3/2}$*):* p^ε_+ *and* p^ε_- *weakly converge in* $L^2(I)$ *and* $L^2(\Omega)$ *to* p^*_+ *and* p^*_-, *satisfying the coupled equation*

$$\begin{aligned} &\int_\omega \frac{h^3}{12\mu} \sum_{i=1}^{2} \frac{\partial p^*_+}{\partial x_i} \frac{\partial \phi}{\partial x_i}\, dx + \int_\Omega \mathbf{K} \nabla p^*_- \nabla \phi \, d\Omega \\ &\quad = -\int_\omega \frac{h}{2} \sum_{i=1}^{2} s_i \frac{\partial \phi}{\partial x_i}\, d\omega - \int_\omega s_3 \phi \, d\omega + \int_{\partial\omega} h(s - \bar{g}) \cdot n \phi \, d\gamma \quad \forall \phi \in H^1_\Sigma. \end{aligned} \tag{4.1}$$

Proof. For any ϕ in H^1_Σ, we extend it to a function Φ defined on I so that $\Phi(x, z) = \phi(x)$ in I and take it as a test function for (2.10). After rescaling we obtain

$$\int_{\Omega^\varepsilon} \mathrm{div} u^{\varepsilon\eta} \Phi \, dX + \eta \int_I \left(\frac{\partial u^{\varepsilon\eta}_1}{\partial x_1} + \frac{\partial u^{\varepsilon\eta}_2}{\partial x_2} \right) \Phi \, dI + \int_I \left(\frac{\partial u^{\varepsilon\eta}_3}{\partial z} \right) \Phi \, dI = 0.$$

Multiplying by η^{-1} and using the Green formula,

$$\begin{aligned} &\int_I \left(u^{\varepsilon\eta}_1 \frac{\partial \phi}{\partial x_1} + u^{\varepsilon\eta}_2 \frac{\partial \phi}{\partial x_2} \right) dI + \frac{\varepsilon}{\eta^{3/2}} \int_\Omega \frac{\sqrt{\eta}}{\varepsilon} u^{\varepsilon\eta} \nabla \Phi \, d\Omega \\ &\quad = \int_{\Gamma_s \cup \Gamma_+} (u^{\varepsilon\eta}_1 n_1 + u^{\varepsilon\eta}_2 n_2) \phi \, d\gamma + \int_\omega s_3 \phi \, d\omega. \end{aligned} \tag{4.2}$$

Propositions 3.2, 3.3, and 3.8 imply that each integral in (4.2) is bounded and has a finite limit. So, coupling effects are obtained through the factor $\varepsilon/\eta^{3/2}$. In the

hypotheses of the theorem, it is equal to $O(1)$. For the sake of simplicity, we take it as equal to 1, so, taking the limit of each term in (4.2), we obtain

$$\int_I \left(u_1^* \frac{\partial \phi}{\partial x_1} + u_2^* \frac{\partial \phi}{\partial x_2}\right) dI + \int_\Omega v^* \nabla \Phi \, d\Omega$$
$$= -\int_\omega \left(s_1 \frac{\partial h}{\partial x_1} + s_2 \frac{\partial h}{\partial x_2} - s_3\right) \phi \, d\omega + \int_{\Gamma_s} g \cdot n \phi \, d\gamma.$$

As $\Phi = \phi(x_1, x_2, 0)$ in I, we take the z-average in I.

$$\int_\omega h \bar{u}^* \cdot \nabla_x \phi \, d\omega + \int_\Omega v^* \cdot \nabla \Phi \, d\Omega = \int_\omega (-s \cdot \nabla_x h + s_3) \phi \, d\omega + \int_{\partial\omega} h \bar{g} \cdot n \phi \, d\gamma. \tag{4.3}$$

Using (3.11) and (3.16) we obtain

$$\int_\omega \frac{h^3}{12\mu} \nabla_x p_+^* \cdot \nabla_x \phi \, d\omega + \int_\Omega \mathbf{K} \nabla p_-^* \cdot \nabla \phi \, d\Omega$$
$$= \int_\omega s \cdot \nabla_x (h\phi) \, d\omega - \int_\omega \frac{h}{2} s \cdot \nabla_x \phi \, d\omega - \int_\omega s_3 \phi \, d\omega - \int_{\partial\omega} h \bar{g} \cdot n \phi \, d\gamma.$$

The proof is completed by using Green's formula on the first term of the right-hand side. □

5. Asymptotic behavior of the pressure on Σ. The main result in this section is the continuity of the pressure on Σ for $\varepsilon = O(\eta^{3/2})$. The global pressure is determined in the whole domain D^ε by (2.9), (2.10) up to an additive constant. In §3 it was extended in Ω by Tartar's technique (see Lemma 3.7). It has been shown that the extension is obtained by defining p^ε in T_ε using a particular constant.

From now on, we call p^ε the unique determination of this extended pressure whose global constant is chosen so that

$$\int_{I^\eta \cup \Omega} p^\varepsilon \, dX = 0, \tag{5.1}$$

and we try to find its behavior when ε tends to 0.

$$p^\varepsilon = p_+^\varepsilon + c_+^\varepsilon \quad \text{in } I, \quad \text{with } c_+^\varepsilon = \frac{1}{|I|} \int_I p^\varepsilon \, dI, \tag{5.2}$$

$$p^\varepsilon = p_-^\varepsilon + c_-^\varepsilon \quad \text{in } \Omega, \quad \text{with } c_-^\varepsilon = \frac{1}{|\Omega|} \int_\Omega p^\varepsilon \, d\Omega. \tag{5.3}$$

We already know the estimates on p_+^ε and p_-^ε, so a first step is to find estimates on the two constants.

5.1. Convergence of the pressure averages.

LEMMA 5.1. *The following estimates are valid:*

$$|c_+^\varepsilon| \le C/\eta^2, \qquad |c_-^\varepsilon| \le C/\eta.$$

Proof. (5.1) gives a relation between the two constants:

$$\eta |I| c_+^\varepsilon + |\Omega| c_-^\varepsilon = 0. \tag{5.4}$$

Let χ be any function in $D(\omega)$; there exist functions ϕ of $H^1(I)$ and φ of $H^1(\Omega)$ such that ϕ and φ are equal to χ on Σ and to 0 on the boundary of $D = \Omega \cup I \cup \Sigma$. We define

$$\psi = (0,0,\varphi) \quad \text{and} \quad \Phi = (0,0,\phi) \quad \text{in } I, \Phi = R^\varepsilon(\psi) \quad \text{in } \Omega^\varepsilon,$$

where R^ε is defined as in Lemma 3.6. Φ is a test function for (2.9). After rescaling, we obtain

$$\mu \int_I \left(\eta^3 \frac{\partial u_3^\varepsilon}{\partial x_1} \frac{\partial \phi}{\partial x_1} + \eta^3 \frac{\partial u_3^\varepsilon}{\partial x_2} \frac{\partial \phi}{\partial x_2} + \eta \frac{\partial u_3^\varepsilon}{\partial z} \frac{\partial \phi}{\partial z} \right) dI + \mu \int_{\Omega^\varepsilon} \eta^2 \nabla u^\varepsilon \nabla R^\varepsilon(\psi)\, dX$$
$$= \int_I \eta^2 (p_+^\varepsilon + c_+^\varepsilon) \frac{\partial \phi}{\partial z}\, dI + \int_{\Omega^\varepsilon} \eta^2 (p_-^\varepsilon + c_-^\varepsilon) \operatorname{div} R^\varepsilon(\psi)\, d\Omega.$$

But $R^\varepsilon(\psi)/\partial\Omega = \psi/\partial\Omega$ (Lemma 3.6). Using (5.4) we find

$$\eta^2 c_+^\varepsilon \left(-1 + \eta \frac{|I|}{|\Omega|} \right) \int_\omega \chi\, d\omega = \mu \int_I \left(\eta^3 \frac{\partial u_3^\varepsilon}{\partial x_1} \frac{\partial \phi}{\partial x_1} + \eta^3 \frac{\partial u_3^\varepsilon}{\partial x_2} \frac{\partial \phi}{\partial x_2} + \eta \frac{\partial u_3^\varepsilon}{\partial z} \frac{\partial \phi}{\partial z} \right) dI \tag{5.5}$$
$$+ \mu \int_{\Omega^\varepsilon} \eta^2 \nabla u^\varepsilon \nabla R^\varepsilon(\psi)\, dX - \int_I \eta^2 p_+^\varepsilon \frac{\partial \phi}{\partial z}\, dI - \int_{\Omega^\varepsilon} \eta^2 p_-^\varepsilon \operatorname{div} R^\varepsilon(\psi)\, d\Omega.$$

Now we will show that all the integrals of the right-hand side are bounded. The first one is bounded by $K\eta$ from (3.2). Using (3.3) and (3.13), we obtain

$$\int_{\Omega^\varepsilon} \eta^2 \nabla u^\varepsilon \nabla R^\varepsilon(\psi)\, dX \le \varepsilon |\nabla R^\varepsilon(\psi)|_\Omega \le K.$$

Moreover, Proposition 3.4 implies

$$\int_I \eta^2 p_+^\varepsilon \frac{\partial \phi}{\partial z}\, dI \le K.$$

Writing $\psi = \varphi e_3$, where e_3 is the z unit vector, $\operatorname{div} R^\varepsilon(\psi)$ is bounded in $L^2(\Omega)$ [21], [14] as well as $\eta^2 p_-^\varepsilon$ ((3.14) and $\varepsilon = O(\eta^{3/2})$). So (5.5) shows that $\eta^2 c_+^\varepsilon$ is bounded and (5.4) shows that ηc_-^ε is bounded. □

PROPOSITION 5.2. *There exists a subsequence of $\eta^2 p^\varepsilon$, with p^ε a solution of* (2.9), (2.10), *and* (5.1), *that weakly converges to p^* so that*

$$p^* = p_+^* + c^* \quad \text{in } I, \tag{5.6}$$

$$p^* = p_-^* \quad \text{in } \Omega. \tag{5.7}$$

Proof. From Lemma 5.1 we already know that

$$\eta^2 c_+^\varepsilon \to c^* \quad \text{and} \quad \eta^2 c_-^\varepsilon \to 0. \tag{5.8}$$

Using (5.2), (5.3) and Proposition 4.4, the result is then obvious. □

5.2. Continuity of the pressure on the interface. We introduce an interior layer B^ε with height ε to take into account the influence of the porous medium in the thin film near the interface Σ. (We note that $\varepsilon = \eta^{3/2}$ implies that for small ε, B^ε is contained in I^η.) Then we apply the so-called energy method to find the limit of the pressure on Σ.

Let us recall that $(w^{i\varepsilon}, \Pi^{i\varepsilon})$ are the classical elementary functions defined by Y-periodicity in Ω^ε from the solution of the local Y-periodic problem (3.20). We recall the classical estimates [21].

LEMMA 5.3. *The following estimates are valid:*

$$|w^{i\varepsilon}|_{L^2_{(\Omega)}} \leq K, \quad \varepsilon|\nabla w^{i\varepsilon}|_{L^2_{(\Omega)}} \leq K, \quad |\Pi^{i\varepsilon}|_{L^2_{(\Omega^\varepsilon)}} \leq K.$$

Now, let us first introduce $w^{i+}(y)$ and $\Pi^{i+}(y)$, which are the unique solutions of the Stokes local elementary problem (up to a constant for Π^{i+}) in the y-variables

$$\begin{aligned}
&\mu \Delta_Y w^{i+} = \nabla_Y \Pi^{i+} \quad \text{in } Y,\\
&\operatorname{div}_Y w^{i+} = 0 \quad \text{in } Y,\\
&w^{i+}(y_1, y_2, 0) = w^i(y_1, y_2, 0),\\
&w^{i+}(y_1, y_2, 1) = \int_0^1 \int_0^1 w^i(y, 0)\, dy_1\, dy_2 = k^i\\
&(w^i \text{ is given by (3.20) and } k^i \text{ is a constant of } R^3),\\
&w^{i+}, \Pi^{i+} \text{ are } y_1, y_2\text{-periodic.}
\end{aligned} \tag{5.9}$$

Boundary conditions for w^{i+} satisfy the condition

$$\int_{\partial Y} w^{i+} \cdot n\, dY = 0,$$

so existence and uniqueness are obtained.

From (w^{i+}, Π^{i+}) we build $(w^{i\varepsilon+}, \Pi^{i\varepsilon+})$ interior layer functions in the x-variables defined in B^ε by translation and homothetic transformation. Now we define global functions in $H^1(D^\varepsilon)^3 \times L^2(D^\varepsilon)$:

$$w^\varepsilon = w^{i\varepsilon} \quad \text{in } \Omega^\varepsilon, \quad w^{i\varepsilon+} \quad \text{in } B^\varepsilon,$$

extended to the whole I^η by the constant value k^i of $w^{i\varepsilon+}$ on $\{x_3 = \varepsilon\}$,

$$q^\varepsilon = \Pi^{i\varepsilon} \quad \text{in } \Omega^\varepsilon, \quad \Pi^{i\varepsilon+} \quad \text{in } B^\varepsilon, \quad 0 \quad \text{elsewhere in } I$$

(the index i does not appear in the global functions w^ε and q^ε since the result in Proposition 5.6 will be independent of i).

Let $B = \{(x, z) \in I, x \in \omega, 0 < z < \sqrt{\eta}\}$ be the rescaled interior layer corresponding to B^ε after the change $z = x_3/\eta$. (In fact, B has a height that depends on η, but we do not use the superscript in the notation for the sake of simplicity.)

Set $\sigma = \{(x, z) \in I, x \in \omega, z = \sqrt{\eta}\}$, the upper boundary of B.

LEMMA 5.4. *The following estimates are valid for* $i = 1, 2$:

$$|w^\varepsilon|_{L^2(B)} \leq K\eta^{1/4}, \qquad \left|\frac{\partial w^\varepsilon}{\partial x_i}\right|_{L^2(B)} \leq K/\eta^{5/4},$$

$$\left|\frac{\partial w^\varepsilon}{\partial z}\right|_{L^2(B)} \leq K/\eta^{1/4}, \qquad |q^\varepsilon|_{L^2(B)} \leq K\eta^{1/4}.$$

Proof. w^{i+} and Π^{i+} given by (5.9) are obviously bounded in $H^1(Y)^3$ and $L^2(Y)$:

$$\int_B (w^\varepsilon)^2\, dx\, dz = \frac{1}{\eta} \int_{B^\varepsilon} (w^\varepsilon)^2\, dX = \frac{1}{\eta\varepsilon^2} \int_Y (w^{i+})^2 \varepsilon^3\, dY \leq \sqrt{\eta} K.$$

The same argument is valid for q^ε:

$$\int_B \left(\frac{\partial w^\varepsilon}{\partial x_i}\right)^2 dx\,dz = \sqrt{\eta}\int_Y \left(\frac{\partial w^{i+}}{\partial x_i}\right)^2 dY = \frac{\sqrt{\eta}}{\varepsilon^2}\int_Y \left(\frac{\partial w^{i+}}{\partial y}\right)^2 dY \le K/\eta^{5/2},$$

$$\int_B \left(\frac{\partial w^\varepsilon}{\partial z}\right)^2 dx\,dz = \sqrt{\eta}\int_Y \left(\frac{\partial w^{i+}}{\partial z}\right)^2 dY = \frac{\sqrt{\eta}}{\eta}\int_Y \left(\frac{\partial w^{i+}}{\partial y}\right)^2 dY \le K/\eta^{1/2}. \qquad \square$$

LEMMA 5.5. *The constant k^i defined in* (5.9) *and the permeability matrix* $\mathbf{K}$ *defined by* (3.21) *are such that*

$$k_3^i = \mathbf{K}_{i3}$$

for the i chosen in the right-hand side of the local Stokes problem (3.20), (5.9).

Proof. Extending w^i by zero in T, we get $\operatorname{div}_Y w^i = 0$ in $Y \Rightarrow \int_{\partial\theta} w^i \cdot n = 0$ for any subdomain θ of Y.

Choosing $\theta = \theta_t = \{Y = (y_1, y_2, y_3), 0 \le y_1, y_2 \le 1, 0 < y_3 < t < 1\}$ and taking the periodic condition into account, we obtain the result that $\int_0^1\int_0^1 w_3^i(y_1, y_2, t)\,dy_1\,dy_2$ does not depend on t and is equal to the constant k^i.

Applying Fubini's theorem to (3.21),

$$\mathbf{K}_{i3} = \int_Y w_3^i(y)\,dY = \int_0^1 k_3^i\,dt = k_3^i. \qquad \square$$

Now we use the energy method with a special test function to gain the continuity of the limit pressure.

PROPOSITION 5.6. *The limit pressure is continuous through Σ in this way:*

$$p_+^* + c^* = p_-^* \quad \text{in } L^2(\omega).$$

Proof. The strong formulation of the problem whose solution is $(w^\varepsilon, q^\varepsilon)$ is

$$\begin{aligned}
&-\mu\Delta_X w^\varepsilon + \frac{1}{\varepsilon}\nabla_X q^\varepsilon = \frac{1}{\varepsilon^2}e_i, \quad \operatorname{div}_X w^\varepsilon = 0 \quad \text{in } \Omega^\varepsilon,\\
&-\mu\Delta_X w^\varepsilon + \frac{1}{\varepsilon}\nabla_X q^\varepsilon = 0, \quad \operatorname{div}_X w^\varepsilon = 0 \quad \text{in } B^\varepsilon,\\
&w^\varepsilon = k = \text{ constant}, \quad q^\varepsilon = 0 \quad \text{in } I^\eta \setminus B^\varepsilon.
\end{aligned}$$

By multiplying these equations by ϕu^ε, where ϕ is any regular function in $D(D)$, integrating, and using Green's formula, we obtain

$$\begin{aligned}
&\mu\int_{D^\varepsilon} \nabla w^\varepsilon \nabla\phi u^\varepsilon\,dX - \int_{D^\varepsilon}\frac{q^\varepsilon}{\varepsilon}\operatorname{div}(\phi u^\varepsilon)\,dX\\
&= \int_\Omega \frac{\phi}{\varepsilon^2}u_i^\varepsilon\,dX - \int_\omega \left[\frac{\partial w^\varepsilon}{\partial x_3}\right]\phi u_3^\varepsilon\,d\omega + \int_\sigma \left[\frac{\partial w^\varepsilon}{\partial x_3}\right]\phi u_3^\varepsilon\,d\sigma + \int_\omega \left[\frac{q^\varepsilon}{\varepsilon}\right]\phi u_3^\varepsilon\,d\omega\\
&\quad - \int_\sigma \left[\frac{q^\varepsilon}{\varepsilon}\right]\phi u^\varepsilon\,d\sigma,
\end{aligned} \tag{5.10}$$

where $[\gamma]$ denotes the jump of γ through either Σ or σ (from I to Ω).

We set ϕw^ε as a test function in (2.9) and we compute the difference with (5.10). We obtain

$$\begin{aligned}
&\mu\int_{D^\varepsilon}\left(\nabla u^\varepsilon\nabla(\phi w^\varepsilon) - \nabla w^\varepsilon\nabla(\phi u^\varepsilon)\right)dX - \int_{D^\varepsilon}\left(p^\varepsilon\operatorname{div}(\phi w^\varepsilon) - \frac{q^\varepsilon}{\varepsilon}\operatorname{div}(\phi u^\varepsilon)\right)dX\\
&= -\int_\Omega \frac{\phi}{\varepsilon^2}u_i^\varepsilon\,dX + \text{ integrals on } \omega\cup\sigma.
\end{aligned}$$

This can be written as

$$\mu A^\varepsilon - F^\varepsilon = C^\varepsilon + E^\varepsilon. \tag{5.11}$$

Computing the limit of each factor, we obtain

$$\begin{aligned}
\text{(a)} \qquad A^\varepsilon = {} & \int_{\Omega^\varepsilon} (\nabla u^\varepsilon w^\varepsilon - \nabla w^\varepsilon u^\varepsilon) \nabla \phi \, dX \\
& + \sum_{k=1}^{3} \sum_{j=1}^{2} \eta \int_I \left(\frac{\partial u_k^\varepsilon}{\partial x_j} w_k^\varepsilon - \frac{\partial w_k^\varepsilon}{\partial x_j} u_k^\varepsilon \right) \frac{\partial \phi}{\partial x_j} \, dI \\
& + \sum_{k=1}^{3} \frac{1}{\eta} \int_I \left(\frac{\partial u_k^\varepsilon}{\partial z} w_k^\varepsilon - \frac{\partial w_k^\varepsilon}{\partial z} u_k^\varepsilon \right) \frac{\partial \phi}{\partial z} \, dI.
\end{aligned}$$

Lemmas 3.1, 5.3, and 5.4 and the fact that w_k^ε is a constant in $I \setminus B$ imply

$$\eta^2 A^\varepsilon \to 0, \tag{5.12}$$

$$\begin{aligned}
\text{(b)} \qquad F^\varepsilon = {} & \int_{\Omega^\varepsilon} \left(p^\varepsilon w^\varepsilon - \frac{q^\varepsilon}{\varepsilon} u^\varepsilon \right) \nabla \phi \, dX + \eta \sum_{k=1}^{2} \int_I \left(p^\varepsilon w_k^\varepsilon - \frac{q^\varepsilon}{\varepsilon} u_k^\varepsilon \right) \frac{\partial \phi}{\partial x_k} \, dI \\
& + \int_I \left(p^\varepsilon w_3^\varepsilon - \frac{q^\varepsilon}{\varepsilon} u_3^\varepsilon \right) \frac{\partial \phi}{\partial z} \, dI, \\
\lim \eta^2 F^\varepsilon = {} & \lim \left(\int_{\Omega^\varepsilon} \eta^2 p^\varepsilon w^\varepsilon \nabla \phi \, dX + \eta^2 \int_I p^\varepsilon w_3^\varepsilon \frac{\partial \phi}{\partial z} \, dI \right).
\end{aligned}$$

In the first integral, $\eta^2 p^\varepsilon = \eta^2 (p_-^\varepsilon + c_-^\varepsilon)$ tends to $\eta^2 p_-^*$ in $L^2(\Omega)$ strongly due to the strong convergence of the pressure in the porous medium and because $\eta^2 c_-^\varepsilon$ tends to 0 from Lemma 5.1; moreover, w_j^ε weakly converges towards its average $\mathbf{K}_{i,j}$ (see (3.21)), so

$$\int_\Omega \eta^2 p^\varepsilon w^\varepsilon \nabla \phi \, dX \to \int_\Omega p_-^* \sum_{j=1}^{3} \mathbf{K}_{i,j} \frac{\partial \phi}{\partial x_j} \, dX.$$

Let us now consider the second integral. Lemmas 5.1 and 5.4 and Proposition 3.4 imply

$$\eta^2 \int_B p^\varepsilon w_3^\varepsilon \frac{\partial \phi}{\partial z} \, dI \to 0,$$

so that

$$\begin{aligned}
\lim \int_I \eta^2 p^\varepsilon w_3^\varepsilon \frac{\partial \phi}{\partial z} \, dI &= \lim \int_{I \setminus B} \eta^2 p^\varepsilon w_3^\varepsilon \frac{\partial \phi}{\partial z} \, dI \\
&= \lim \int_{I \setminus B} \eta^2 p^\varepsilon k_3^i \frac{\partial \phi}{\partial z} \, dI = \int_{I \setminus B} (p_+^* + c_+^*) k_3^i \frac{\partial \phi}{\partial z} \, dI.
\end{aligned}$$

Then

$$\lim \eta^2 F^\varepsilon = \int_\Omega p_-^* \sum_{j=1}^{3} \mathbf{K}_{i,j} \frac{\partial \phi}{\partial x_j} \, dX + \int_I (p_+^* + c^*) k_3^i \frac{\partial \phi}{\partial z} \, dI. \tag{5.13}$$

(c) Proposition 3.8 implies

$$\lim \eta^2 C^\varepsilon = -\int_\Omega \phi v_i^* \, dI. \tag{5.14}$$

(d) In $\eta^2 E^\varepsilon$, the integrals on ω contain $\partial w_k^\varepsilon/\partial y_j$ or q^ε, which are regular, bounded, and periodic functions of $L^2(\omega)$ due to (3.22). From Lemma 4.1, u_k^ε tends strongly in $L^2(\omega)$ to zero, so these integrals vanish at the limit. For the integral on σ, we use the inequality

$$\int_\sigma \psi^2 \, d\sigma \le 2\left[\int_\omega \psi^2 \, d\omega + \sqrt{\eta}\int_B \left(\frac{\partial \psi}{\partial z}\right)^2 dx\, dz\right], \tag{5.15}$$

which is easily obtained by a density argument for any ψ in $H^1(B)$. Applying (5.15) to u_k^ε, using Lemmas 3.1 and 4.1, we obtain

$$|u_k^\varepsilon|_{L^2(\sigma)} \le C.$$

The limit of the integrals on σ is computed in the same way as those on ω, and this limit is equal to zero, so

$$\eta^2 E^\varepsilon \to 0. \tag{5.16}$$

Now we are able to find the limit in (5.11), using (5.12)–(5.16):

$$\int_\Omega p_-^* \sum_{j=1}^{3} \mathbf{K}_{i,j} \frac{\partial \phi}{\partial x_j} \, dX - \int_\Omega \phi v_i^* \, dX + \int_I (p_+^* + c^*) k_3^i \frac{\partial \phi}{\partial z} \, dI = 0.$$

Using (3.16) and the fact that p_+^* does not rely on z,

$$-\int_\omega p_-^* \mathbf{K}_{i,3} \phi \, d\omega + \int_\omega (p_+^* + c^*) k_3^i \phi \, d\omega = 0 \quad \forall \phi \in D(D).$$

Lemma 5.5 $\Rightarrow p_-^* = p_+^* + c^*$ in $D'(\omega)$. This is the continuity of the pressure $\eta^2 p^\varepsilon$ at the limit.

6. The limit problem. We now write the main result of the paper in the critical case $\varepsilon = O(\eta^{3/2})$.

THEOREM 6.1. *The unique solution of problem* (2.9), (2.10) *is such that* $\eta^2 p^\varepsilon$ *weakly converges in* $L^2(\Omega)$ *to* p^*, *which is the unique solution in* $H^1_\Sigma(\Omega)$ *of*

$$\begin{aligned} &\int_\omega \frac{h^3}{12\mu} \sum_{i=1}^{2} \frac{\partial p^*}{\partial x_i} \frac{\partial \phi}{\partial x_i} d\omega + \int_\Omega \mathbf{K} \nabla p^* \nabla \phi \, d\Omega = \int_{\partial\omega} h(s - \bar{g}) \cdot n \phi \, d\sigma \\ &\quad - \int_\omega \left(\frac{h}{2} \sum_{i=1}^{2} s_i \frac{\partial \phi}{\partial x_i} + s_3 \phi\right) d\omega \quad \forall \phi \in H^1_\Sigma(\Omega), \end{aligned} \tag{6.1}$$

$$\int_\Omega p^* \, dX = 0. \tag{6.2}$$

Proof. Using Propositions 4.4, 5.2, and 5.6, we find that there exists a subsequence of p^ε that converges to p^*, which is the solution of (6.1), (6.2) as it belongs to $H^1_\Sigma(\Omega)$. The uniqueness is obvious, so the whole sequence converges. □

The limit velocities are obvious to compute, as shown in the following theorem.

THEOREM 6.2. *In the rescaled thin film, the limit velocity is*

$$u_i^* = \frac{\partial p^*}{\partial x_i} z(z-h) + s_i\ z/h, \quad i = 1,2, \quad u_3^* = 0; \tag{6.3}$$

in the porous medium, u^ε/η weakly converges in $L^2(\Omega)$ to

$$v^* = -\mathbf{K}\nabla p^*, \tag{6.4}$$

where p^* is the unique solution of problem (6.1), (6.2).

7. Concluding remarks. We completely determined the limit problem in the critical case $\varepsilon = O(\eta^{3/2})$, that is, the first term of an asymptotic expansion of the velocity and the pressure. Equality (4.2) led us to the choice of the critical values of the small parameters. We return to (4.2) to see what conclusion can be given when $\varepsilon \neq O(\eta^{3/2})$.

—If $\varepsilon < O(\eta^{3/2})$, (4.2) implies that the asymptotic analysis is obtained in the thin film without any influence from the porous medium. The "limit pressure" p^* is the solution of the Reynolds equation (see (3.5)) with $t_3^* = 0$, and the coupling will take place only if we want to find the second-order terms. We could say the porous medium is too compact for the free flow to penetrate inside.

—If $\varepsilon > O(\eta^{3/2})$, (4.2) implies, after multiplication by $\eta^{3/2}/\varepsilon$, that the porous medium has a trivial asymptotic solution, zero velocity, and constant pressure. The velocity of the free flow cannot influence the porous medium due to the lateral boundary layer, but Corollary 4.2 is not always true and a more precise study should be carried out to give the complete coupling at the second order.

REFERENCES

[1] G. BAYADA AND M. CHAMBAT, *The transition between the Stokes equations and the Reynolds equation: A mathematical proof*, Appl. Math. Opt., 14 (1986), pp. 73–93.

[2] G. S. BEAVERS AND D. D. JOSEPH, *Boundary conditions at a naturally permeable wall*, J. Fluid Mech., 30 (1967), pp. 197–207.

[3] M. BOUKROUCHE AND G. BAYADA, *Mathematical model, existence and uniqueness of a cavitation problem in a porous journal bearing*, Nonlinear Anal., 20 (1993), pp. 895–920.

[4] A. BOURGEAT, H. ELAMRI, AND R. TAPIERO, *Existence d'une taille critique pour une fissure dans un milieu poreux*, in Second Colloque Franco Chilien de Mathématiques Appliquées, Cepadués Edts., Toulouse, France, 1991, pp. 67–80.

[5] A. BOURGEAT AND R. TAPIERO, *Homogéneisation dans un domaine perforé incluant une couche pleine de faible épaisseur*, Preprint 126, Ura740/Cnrs Lyon, 1991, pp. 1–15.

[6] G. CIMATTI, *On the mathematical theory of porous journal bearing*, Meccanica, 15 (1980), pp. 112–117.

[7] T. DUMONT, *Décomposition par projection de certains problèmes aux limites elliptiques non linéaires*, Ph.D. Thesis, Université de Lyon, 1978.

[8] H. I. ENE AND E. SANCHEZ-PALENCIA, *Equations et phénomènes de surface pour l'écoulement dans un modèle de milieu poreux*, J. Mécanique, 14 (1975), pp. 73–107.

[9] D. D. JOSEPH AND L. N. TAO, *Lubrication of a porous journal bearing Stokes solution*, J. Appl. Mech., Trans. A.S.M.E., 33 (1966), pp. 753–760.

[10] K. LEMRABET, *Problème de Ventcel pour le système de l'élasticité dans un domaine de R^3*, C.R. Acad. Sci. Paris Sér. I, Math., 304 (1987), pp. 151–154.

[11] T. LEVY, *Filtration in a porous fissured rock: Influence of the fissures connexity*, European J. Mech., B, 9 (1990), pp. 309–327.

[12] T. LEVY AND E. SANCHEZ-PALENCIA, *On boundary conditions for fluid flow in porous media*, Internat. J. Engrg. Sci., 13 (1975), pp. 923–940.

[13] M. H. MEURISSE AND B. GUIDICELLI, *On the mathematical models of porous journal bearing*, in Mathematical Modelling in Lubrication, Publ. Univ. Vigo, Spain, 1991, pp. 79–85.

[14] A. MIKELIC AND I. AGANOVIC, *Homogenization in a porous medium under a nonhomogeneous boundary condition*, Boll. Un. Mat. Ital., 7 (1987), pp. 171–180.

[15] G. NGUETSENG, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[16] B. R. REASON AND A. H. SIEW, *A refined numerical solution for the hydrodynamic lubrication of finite porous journal bearing*, Proc. Inst. Mech. Engrs, 199 (1985), pp. 85–93.

[17] S. RICHARDSON, *A model for the boundary condition of a porous material, Part* 2, J. Fluid Mech., 49 (1971), pp. 327–336.

[18] W. T. ROULEAU, *Hydrodynamic lubrication of narrow press fitted porous metal bearing*, J. Basic. Engrg., Trans. A.S.M.E., D (1963), pp. 123–128.

[19] W. T. ROULEAU AND L. I. STEINER, *Hydrodynamic porous journal bearings; Part* 1, *Finite full journal bearings*, J. Lub. Tech., Trans. A.S.M.E., 96 (1974), pp. 346–353.

[20] P. G. SAFFMAN, *On the boundary condition at the surface of a porous medium*, Stud. Appl. Math., L(2) (1971), pp. 93–101.

[21] E. SANCHEZ-PALENCIA, *Nonhomogeneous media and vibration theory*, Lecture Notes in Phys. 127, Springer-Verlag, New York, 1980.

[22] R. TEMAM, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1979.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1130–1142, September 1995

003

EXISTENCE OF SHOCK PROFILES FOR VISCOELASTIC MATERIALS WITH MEMORY*

SHUICHI KAWASHIMA† AND HARUMI HATTORI‡

Abstract. In this paper we discuss the necessary and sufficient conditions for the existence of smooth monotone shock profiles for viscoelastic materials with memory. We also discuss the uniqueness. We consider both convex and nonconvex constitutive relations. In the case of nonconvex constitutive relations, we include a degenerate case where the speed of the shock profile is equal to the speed of the equilibrium characteristics at one of the end states. This was not discussed in previous literature.

Key words. hyperbolic system, shock profile, viscoelasticity

AMS subject classifications. Primary, 35L60, 70A10; Secondary, 35L65

1. Introduction.

In this paper, we shall dicuss the existence of smooth shock profiles for a system

$$\begin{aligned} v_t &= u_x, \\ u_t &= \sigma(v)_x + \int_{-\infty}^{t} a'(t-\tau)\eta(v)_x \, d\tau \end{aligned} \tag{1.1}$$

arising in viscoelastic materials with fading memory. In (1.1), v and u are strain and velocity, respectively. In this paper we make the following assumptions on σ, η, and $a(t)$.

Assumption 1.1. (i) σ and η are smooth and $\sigma' > 0$, $\eta' > 0$. We also require that

$$\chi(v) \equiv \sigma(v) - a(0)\eta(v) \tag{1.2}$$

satisfies

$$\chi' > 0. \tag{1.3}$$

(ii) For $a(t)$ we assume

$$\begin{aligned} &a, a', a'' \in L^1(0,\infty), \\ &a(t) > 0, \quad a'(t) < 0, \qquad t \in [0,\infty). \end{aligned} \tag{1.4}$$

We call $\pm\sqrt{\sigma'}$ and $\pm\sqrt{\chi'}$ the instantaneous and equlibrium characteristic speeds, respectively. The shock profiles are defined as follows.

DEFINITION 1.1. *The function $(V,U)(x-st)$ is the shock profile for* (1.1) *connecting constant states (v_-,u_-) and (v_+,u_+) if and only if the following hold:*

(i) *$(V,U)(\xi)$ is smooth and strictly monotone as a function of $\xi \in \mathcal{R}$, and*

$$(V,U)(\xi) \to (v_\pm, u_\pm) \quad \text{as} \quad \xi \to \pm\infty. \tag{1.5}$$

* Received by the editors May 19, 1993; accepted for publication January 3, 1994.

† Department of Applied Science, Faculty of Engineering 36, Kyushu University, Fukuoka 812, Japan.

‡ Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506. The research of this author was supported by U.S. Army grant DAAL 03-89-G-0088 and DEPSCOR grant DAAH04-94-G-0246.

(ii) $(V,U)(x-st)$ *satisfies* (1.1).

Here, s is the speed of the shock profile.

When $s \gtrless 0$, for the behavior of $V(\xi)$ as $\xi \to \pm\infty$, we require that

$$w_\pm(\tau) \equiv \lim_{\xi\to\pm\infty} \frac{V(\xi+s\tau)-v_\pm}{V(\xi)-v_\pm} > 0 \quad \text{for } \tau \geq 0. \tag{1.6}$$

Note that as long as the limits $w_\pm(\tau)$ exist, we always have $0 \leq w_\pm(\tau) \leq 1$ by the monotonicity of $V(\xi)$.

Our goal is to obtain the necessary and sufficient conditions for the existence of the shock profiles for (1.1) satisfying (1.6), and also to discuss the uniqueness. We shall include the case where $\chi'(v_\pm) = s^2$ for $s \gtrless 0$. This is precisely the case which was not covered in previous literature. The above conditions consist of the following three conditions. First, the end states $(v_\pm, u_\pm)$ satisfy the Rankine–Hugoniot condition

$$\begin{aligned} -c(v_+ - v_-) &= u_+ - u_-, \\ -c(u_+ - u_-) &= \chi_+ - \chi_-, \end{aligned} \tag{1.7}$$

where $\chi_\pm = \chi(v_\pm)$. We shall use this type of abbreviation throughout the paper. Second, the following entropy condition is satisfied: for $s \gtrless 0$,

$$s^2 \lessgtr \frac{\chi(v)-\chi_-}{v-v_-} \quad \text{for } v \in (v_+, v_-). \tag{1.8}$$

Under (1.7) this entropy condition is equivalent to

$$s^2 \gtrless \frac{\chi(v)-\chi_+}{v-v_+} \quad \text{for } v \in (v_+, v_-). \tag{1.9}$$

Third, the following nonresonance condition holds:

$$s^2 < \sigma'(v) \quad \text{for } v \in [v_+, v_-]. \tag{1.10}$$

We should note here that (a,b) or $[a,b]$ in this paper does not necessarily imply $a \leq b$. This will be assumed throughout the paper.

System (1.1) describes a one-dimensional motion of an unbounded, homogeneous, viscoelastic bar. The integral term represents the memory effect of the material. Regarding the existence of shock profiles for (1.1), Pipkin [6] discussed the case where the kernel $a(t)$ is an exponential function. Greenberg [1], [2] and Greenberg and Hasting [3] disucussed this in the case where both σ and χ are concave. Liu [5] discussed the existence of shock profiles for (1.1) without assuming the convexity of σ and χ. It is interesting to observe that the shock profile may become discontinuous. The stability of shock profiles for (1.1) with additional assumptions on the kernel $a(t)$ was discussed in [4].

This paper consists of four sections. In §2 we discuss the necessary conditions for the existence of shock profiles and in §3 we give the sufficient conditions. In §4 we discuss the uniqueness of shock profiles. We are interested in the case where χ is nonconvex and $s^2 = \chi'(v_\pm)$ for $s \gtrless 0$. The proofs in §§3 and 4 basically follow Greenberg and Hasting [3].

2. Necessary conditions. In this section we shall prove the following theorem stating the necessary conditions for the existence of the shock profiles for (1.1) satisfying (1.6).

THEOREM 2.1. *Suppose Assumption* 1.1 *is satisfied. If there exists a shock profile* $(V,U)(x-st)$ *for* (1.1) *connecting* (v_-,u_-) *and* (v_+,u_+), *then conditions* (1.7), (1.8), *and*

$$s^2 < \sigma'(v) \quad \text{for } v \in (v_+, v_-) \tag{2.1}$$

hold. Furthermore, if there exists a shock profile satisfying (1.6), *the stronger nonresonance condition* (1.10) *is satisfied.*

The Rankine–Hugoniot condition. Suppose $(V,U)(x-st)$ is a shock profile for (1.1) connecting (v_-,u_-) and (v_+,u_+); then we see

$$\begin{aligned} &-sV' - U' = 0, \\ &-sU' - \left\{\sigma(V) + \int_0^\infty a'(\tau)\eta(V(\xi+s\tau))\,d\tau\right\}' = 0. \end{aligned} \tag{2.2}$$

Integrating the above relations, we have

$$\begin{aligned} &-sV - U = A_1, \\ &-sU - \left\{\sigma(V) + \int_0^\infty a'(\tau)\eta(V(\xi+s\tau))\,d\tau\right\} = A_2, \end{aligned} \tag{2.3}$$

where A_1, A_2 are constants of integration. From $(V,U)(\pm\infty) = (v_\pm, u_\pm)$ and

$$\sigma(v_\pm) + \int_0^\infty a'(\tau)\eta(v_\pm)d\tau = \sigma_\pm - a(0)\eta_\pm = \chi_\pm,$$

we see

$$A_1 = -sv_\pm - u_\pm, \quad A_2 = -su_\pm - \chi_\pm. \tag{2.4}$$

This implies the Rankine–Hugoniot condition (1.7).

The entropy condition. Solving the first equation in (2.3) for U and substituting in the second equation, we have

$$-s^2V + \sigma(V) = -\int_0^\infty a'(\tau)\eta(V)(\xi+s\tau)\,d\tau + A, \tag{2.5}$$

where

$$A = sA_1 - A_2 = -s^2v_\pm + \chi_\pm. \tag{2.6}$$

The integration of the memory term in (2.5) yields

$$s\int_0^\infty a(\tau)(\eta'(V)V')(\xi+s\tau)\,d\tau = -s^2V + \chi(V) - A. \tag{2.7}$$

Consider the case where $s > 0$ and $v_+ \lesseqgtr v_-$. Because $V(\xi)$ is strictly monotone, $V'(\xi) \lesseqgtr 0$ for $\xi \in \mathcal{R}$. Therefore, from Assumption 1.1, we see that the left-hand side (LHS) of (2.7) $\lesseqgtr 0$ for $\xi \in \mathcal{R}$. This implies

$$-s^2v + \chi(v) - A \lesseqgtr 0 \quad \text{for } v \in (v_+, v_-).$$

So using (2.6), we obtain

$$(v - v_-)\left(-s^2 + \frac{\chi(v) - \chi_-}{v - v_-}\right) \lesseqgtr 0 \quad \text{for } v \in (v_+, v_-).$$

This yields

$$s^2 < \frac{\chi(v) - \chi_-}{v - v_-} \quad \text{for } v \in (v_+, v_-),$$

which is the entropy condition for $s > 0$. The case where $s < 0$ is similarly shown.

The nonresonance condition. Differentiating (2.5) with respect to ξ, we find

$$(-s^2 + \sigma'(V))V' = -\int_0^\infty a'(\tau)(\eta'(V)V')(\xi + s\tau)\,d\tau. \tag{2.8}$$

If $v_+ \lesseqgtr v_-$, then $V'(\xi) \lesseqgtr 0$ for $\xi \in \mathcal{R}$. This fact, along with Assumption 1.1, implies that the right-hand side (RHS) of (2.8) $\lesseqgtr 0$ for $\xi \in \mathcal{R}$ and, therefore,

$$-s^2 + \sigma'(v) > 0 \quad \text{for } v \in (v_+, v_-).$$

We now show that if $V(\xi)$ satisfies (1.6), then

$$s^2 < \sigma'(v_\pm) \tag{2.9}$$

is also satisfied. Taking the limit as $v \to v_-$ or v_+ in (1.8) and (1.9), we see

$$\begin{aligned} &\chi'(v_+) \le s^2 \le \chi'(v_-) \quad \text{if } s > 0, \\ &\chi'(v_-) \le s^2 \le \chi'(v_+) \quad \text{if } s < 0. \end{aligned} \tag{2.10}$$

Combining this with $\chi' < \sigma'$, we obtain

$$s^2 < \sigma'(v_\mp) \quad \text{for } s \gtrless 0.$$

Next, we show

$$s^2 < \sigma'(v_\pm) \quad \text{for } s \gtrless 0.$$

We discuss the case where $s > 0$. Expanding (2.5) about v_+, we see

$$\begin{aligned} &-s^2 v_+ + \sigma(v_+) + (-s^2 + \sigma'(v_+))(V(\xi) - v_+) + O(|V(\xi) - v_+|^2) \\ &\quad = A - \int_0^\infty a'(\tau)\{\eta(v_+) + \eta'(v_+)(V(\xi + s\tau) - v_+) + O(|V(\xi + s\tau) - v_+|^2)\}\,d\tau \end{aligned} \tag{2.11}$$

as $\xi \to +\infty$. Using

$$A - \int_0^\infty a'(\tau)\eta(v_+)d\tau = -s^2 v_+ + \sigma(v_+)$$

and dividing (2.11) by $(V(\xi) - v_+)$, we obtain

$$\begin{aligned} &-s^2 + \sigma'(v_+) + O(|V(\xi) - v_+|) \\ &\quad = -\int_0^\infty a'(\tau)\left\{\eta'(v_+)\frac{V(\xi + s\tau) - v_+}{V(\xi) - v_+} + O(|V(\xi + s\tau) - v_+|)\right\} d\tau \end{aligned} \tag{2.12}$$

as $\xi \to +\infty$. Taking the limit as $\xi \to \infty$, we see

$$-s^2 + \sigma'(v_+) = -\int_0^\infty a'(\tau)\eta'(v_+)w_+(\tau)d\tau > 0. \tag{2.13}$$

So, if (1.6) is satisfied, then

$$s^2 < \sigma'(v_+).$$

The case where $s < 0$ is proved in the same manner. □

3. Sufficient conditions. We shall discuss the sufficient conditions for the existence of the shock profiles. Because of (2.10), we discuss the following two cases.

Case A.

$$\chi'(v_+) < s^2 \quad \text{if } s > 0, \qquad \chi'(v_-) < s^2 \quad \text{if } s < 0, \tag{3.1}$$

Case B.

$$\begin{aligned} &\chi'(v_+) = s^2 \quad \text{and} \quad \chi''(v_+) \neq 0 \quad \text{if } s > 0, \\ &\chi'(v_-) = s^2 \quad \text{and} \quad \chi''(v_-) \neq 0 \quad \text{if } s < 0. \end{aligned} \tag{3.2}$$

In Case B we see from (1.8) and (3.2) that

$$\begin{aligned} &\chi''(v_+) \lesseqgtr 0 \quad \text{for } v_+ \lesseqgtr v_- \quad \text{if } s > 0, \\ &\chi''(v_-) \gtreqless 0 \quad \text{for } v_+ \lesseqgtr v_- \quad \text{if } s < 0. \end{aligned} \tag{3.3}$$

This can be seen as follows. When $s > 0$, expanding (1.9) in a Taylor series about $v = v_+$, we obtain

$$s^2 > \chi'(v_+) + \frac{1}{2}\chi''(v_+)(v - v_+) + O(|v - v_+|^2).$$

Since $v_+ \lesseqgtr v$ for $v \in (v_+, v_-)$ if $v_+ \lesseqgtr v_-$, we obtain (3.3) from (3.2). The case where $s < 0$ can be shown in a similar way.

Case A is a nondegenerate case. This case has been discussed in [3] and [5]. If χ is strictly convex (or concave), we have only Case A. Case B is one of the simplest degenerate cases and is limited to the case where χ has an inflection point. We shall not discuss the cases where χ has higher degeneracy. For example, we do not discuss the case where $s > 0$ and

$$\chi'(v_+) = s^2, \quad \chi^{(j)}(v_+) = 0, \quad 2 \le j \le m, \quad \chi^{(m+1)}(v_+) \neq 0 \tag{3.4}$$

for an integer $m \ge 2$.

THEOREM 3.1. *Suppose Assumption* 1.1 *is satisfied. Assume that* $(v_\pm, u_\pm)$ *and* s *satisfy* (1.7), (1.8), *and* (1.10) *and that* (3.1) *or* (3.2) *holds. In the case where* (3.2) *holds, we also assume that the kernel* $a(t)$ *satisfies*

$$(1+t)^3 a'(t) \in L^1(0, \infty). \tag{3.5}$$

Then, there exists a shock profile for (1.1) *which connects* (v_-, u_-) *and* (v_+, u_+), *and satisfies* (1.6).

We assume that χ satisfies one of the following:

$$\begin{aligned} &\text{(i)} && \chi''(v) > 0 && \text{for all } v, \\ &\text{(ii)} && \chi''(v) < 0 && \text{for all } v, \\ &\text{(iii)} && \chi''(v) \gtrless 0 && \text{for } v \gtrless 0, \\ &\text{(iv)} && \chi''(v) \lesseqgtr 0, && \text{for } v \gtrless 0. \end{aligned} \tag{3.6}$$

If we assume (1.8), only Case A occurs in (i), (ii), and (iv) of (3.6). In (iii) of (3.6), only Case A or Case B is possible and the higher degeneracies such as those in (3.4) do not occur. Therefore, from Theorems 2.1 and 3.1 we have the following corollary.

COROLLARY 3.1. *Suppose that Assumption* 1.1 *holds. Assume that* χ *satisfies one of the conditions in* (3.6) *and that* (3.5) *also holds in* (iii) *of* (3.6). *Then, there exists a shock profile for* (1.1) *which connects* (v_-, u_-) *and* (v_+, u_+), *and satisfies* (1.6) *if and only if* (1.7), (1.8), *and* (1.10) *are satisfied.*

We prove Theorem 3.1 for $s > 0$ only. Case A is essentially from Greenberg and Hasting [3] and Liu [5]. We follow [3]. Define a function space X as follows: if $v_- > v_+$, then

$$\begin{aligned} X = \{\phi(\xi) \mid \phi(\xi) &\text{ is piecewise continuous, nonincreasing, and} \\ &\phi(\infty) = v_+, \phi(-\infty) \le v_-\}, \end{aligned} \tag{3.7}$$

and if $v_- < v_+$, then

$$\begin{aligned} X = \{\phi(\xi) \mid \phi(\xi) &\text{ is piecewise continuous, nondecreasing, and} \\ &\phi(\infty) = v_+, \phi(-\infty) \ge v_-\}. \end{aligned} \tag{3.8}$$

We also define

$$\rho(v) \equiv -s^2 v + \sigma(v), \quad v \in [v_+, v_-], \tag{3.9}$$

$$K[\phi](\xi) \equiv -\int_0^\infty a'(\tau)\eta(\phi)(\xi + s\tau)d\tau + A, \quad \phi \in X. \tag{3.10}$$

From (1.10) we see that $w \equiv \rho(v)$ is a smooth and monotonically increasing function mapping $[v_+, v_-]$ onto $[\rho_+, \rho_-]$, where $\rho_\pm = \rho(v_\pm)$. So, there exists an inverse function $v = \rho^{-1}(w)$ on $w \in [\rho_+, \rho_-]$, which is smooth and monotonically increasing. Note that

$$K[\phi](\xi) \in [\rho_+, \rho_-] \quad \text{for } \xi \in \mathcal{R} \quad \text{if } \phi \in X.$$

This implies that the operator T defined by

$$T[\phi](\xi) \equiv \rho^{-1}(K[\phi])(\xi), \quad \phi \in X \tag{3.11}$$

is well defined. Now we see (2.5) is written as

$$\rho[V] = K[V], \quad \text{i.e., } V = T[V].$$

Therefore, the existence of a shock profile for (1.1) is equivalent to finding a fixed point of the operator T. In the following lemmas, we discuss the properties of the operator T.

LEMMA 3.1 ([3]). *Assume that* $s > 0$ *and* (1.7) *and* (1.10) *hold. Then,* T *satisfies the following:*

(i) *If* $\phi \in X$, *then* $T[\phi] \in X \cap \mathcal{B}^0$, *and if* $\phi \in X \cap \mathcal{B}^k$, *then* $T[\phi] \in X \cap \mathcal{B}^{k+1}$ ($k \geq 0$, *integer*).

(ii) *If* $\phi_1, \phi_2 \in X$ *and* $\phi_1 \leq \phi_2$, *then* $T[\phi_1] \leq T[\phi_2]$.

In this lemma $\mathcal{B}^k$ means the space of k-times continuously differentiable functions with bounded derivatives on $\mathcal{R}$. We omit the proof. See Lemma 1 in [3].

In Lemmas 3.2 and 3.3 we show the asymptotic behavior of the shock profiles as ϕ approaches one of the end states. Lemmas 3.2 and 3.3 correspond to Cases A and B, respectively.

LEMMA 3.2 ([3]). *Assume that* $s > 0$ *and that* (1.7) *and* (1.10) *hold. Then, in Case* A *there exist nontrivial* $\phi_1, \phi_2 \in X$ *satisfying*

$$\phi_1 < \phi_2, \quad \phi_1 \leq T[\phi_1], \quad T[\phi_2] \leq \phi_2 \quad \text{for} \quad \xi \in \mathcal{R}. \tag{3.12}$$

Proof. Although the proof is given in [3], we show it briefly for the sake of comparison with Lemma 3.3. Set

$$\phi(\xi) = v_+ + \delta e^{-\lambda\xi} + \mu e^{-2\lambda\xi}, \quad \delta = v_- - v_+, \tag{3.13}$$

where $\lambda > 0$ and $\mu \in \mathcal{R}$ are constants to be determined. After expanding $\rho(\phi)(\xi)$ and $K[\phi](\xi)$ in powers of $e^{-\lambda\xi}$, we see, as $\xi \to +\infty$, that

$$\begin{aligned}
&\rho(\phi)(\xi) - K[\phi](\xi) \\
&\quad = \delta\left\{-s^2 + \sigma'(v_+) + \eta'(v_+)\int_0^\infty a'(\tau)e^{-\lambda s\tau}d\tau\right\}e^{-\lambda\xi} \\
&\quad\quad + \left\{\mu\left(-s^2 + \sigma'(v_+) + \eta'(v_+)\int_0^\infty a'(\tau)e^{-2\lambda s\tau}d\tau\right)\right. \\
&\quad\quad\quad \left. + \frac{1}{2}\delta^2\left(\sigma''(v_+) + \eta''(v_+)\int_0^\infty a'(\tau)e^{-2\lambda s\tau}d\tau\right)\right\}e^{-2\lambda\xi} + O(e^{-3\lambda\xi}) \\
&\quad \equiv \delta C_1(\lambda)e^{-\lambda\xi} + C_2(\lambda,\mu)e^{-2\lambda\xi} + O(e^{-3\lambda\xi}).
\end{aligned} \tag{3.14}$$

The following statements are shown in [3]:

1. There exists a unique $\lambda > 0$ such that $C_1(\lambda) = 0$.
2. There exists μ_+ and μ_- such that $\mu_+ > \mu_-$ and $C_2(\lambda, \mu_+) > 0 > C_2(\lambda, \mu_-)$.

Using the above λ and $\mu_\pm$, for $v_+ < v_-$ we set

$$\phi_1(\xi) = \begin{cases} v_+ + \delta e^{-\lambda\xi} + \mu_- e^{-2\lambda\xi}, & \xi \geq R, \\ \phi_1(R), & \xi < R, \end{cases} \tag{3.15}$$

$$\phi_2(\xi) = \begin{cases} v_+ + \delta e^{-\lambda\xi} + \mu_+ e^{-2\lambda\xi}, & \xi \geq R, \\ v_-, & \xi < R, \end{cases} \tag{3.16}$$

where R is a large positive number. This R is chosen so that

$$\rho(\phi_1)(\xi) \leq K[\phi_1](\xi), \quad \rho(\phi_2)(\xi) \geq K[\phi_2](\xi)$$

are satisfied for $\xi \geq R$. This implies that, for $\xi \geq R$,

$$\phi_1(\xi) \leq T[\phi_1](\xi), \quad \phi_2(\xi) \geq T[\phi_2](\xi). \tag{3.17}$$

It is shown in [3] that (3.17) actually holds for $\xi < R$ as well.

For $v_+ > v_-$ we set

$$\phi_1(\xi) = \begin{cases} v_+ + \delta e^{-\lambda\xi} + \mu_- e^{-2\lambda\xi}, & \xi \geq R, \\ v_-, & \xi < R, \end{cases} \tag{3.18}$$

$$\phi_2(\xi) = \begin{cases} v_+ + \delta e^{-\lambda\xi} + \mu_+ e^{-2\lambda\xi}, & \xi \geq R, \\ \phi_2(R), & \xi < R. \end{cases} \tag{3.19}$$

The rest of the proof is the same as the case where $v_- > v_+$. □

LEMMA 3.3. *Assume that* $s > 0$ *and* (1.7), (1.8), *and* (1.10) *hold. We also assume* (3.5). *Then, in Case* B *there exist nontrivial* $\phi_1, \phi_2 \in X$ *satisfying* (3.12).

Proof. Set

$$\phi(\xi) = v_+ + (\alpha + \beta\zeta^{-1} + \gamma\zeta^{-2})\xi^{-1}, \quad \zeta^{-1} = \xi^{-1}\log\xi, \tag{3.20}$$

where α, β, and γ are constants to be determined. For a given smooth function f, we have the following expansions for large ξ:

$$\begin{aligned} f(\phi)(\xi) = {} & f(v_+) + f'(v_+)(\alpha + \beta\zeta^{-1} + \gamma\zeta^{-2})\xi^{-1} \\ & + \frac{1}{2}f''(v_+)(\alpha^2 + 2\alpha\beta\zeta^{-1} + (2\alpha\gamma + \beta^2)\zeta^{-2})\xi^{-2} \\ & + \frac{1}{6}f'''(v_+)\alpha^3\xi^{-3} + O(\xi^{-4}\log\xi), \end{aligned} \tag{3.21}$$

(3.22)

$$\begin{aligned} f(\phi)(\xi + s\tau) = {} & f(v_+) + f'(v_+)(\alpha + \beta\zeta^{-1} + \gamma\zeta^{-2})\xi^{-1} \\ & + \left(\frac{1}{2}f''(v_+)\alpha - s\tau f'(v_+)\right)(\alpha + \beta\zeta^{-1})\xi^{-2} \\ & + \left(\frac{1}{6}f'''(v_+)\alpha^3 - s\tau f''(v_+)\alpha^2 + (s\tau)^2 f'(v_+)\alpha + s\tau f'(v_+)\beta\right)\xi^{-3} \\ & + \left\{(f''(v_+)\alpha - 3s\tau f'(v_+))\gamma + \frac{1}{2}ff''(v_+)\beta^2\right\}\zeta^{-2}\xi^{-2} + O(\tau^3\xi^{-4}\log\xi), \end{aligned}$$

where in (3.22) the order O is uniform in τ large. Using (3.21) for $\rho(v)$ and (3.22) for $\eta(v)$, we obtain for large ξ,

$$\begin{aligned} & \rho(\phi)(\xi) - K[\phi](\xi) \\ & = (-s^2 + \chi'(v_+))(\alpha + \beta\zeta^{-1} + \gamma\zeta^{-2})\xi^{-1} \\ & \quad + \left(\frac{1}{2}\chi''(v_+)\alpha + sa_1\eta'(v_+)\right)(\alpha + \beta\zeta^{-1})\xi^{-2} \\ & \quad + \left(\frac{1}{6}\chi'''(v_+)\alpha^3 + sa_1\eta''(v_+)\alpha^2 - s^2a_2\eta'(v_+)\alpha - sa_1\eta'(v_+)\beta\right)\xi^{-3} \\ & \quad + \left\{(\chi''(v_+)\alpha + 3sa_1\eta'(v_+))\gamma + \frac{1}{2}\chi''(v_+)\beta^2\right\}\zeta^{-2}\xi^{-2} + O(\xi^{-4}\log\xi) \\ & \equiv C_o \cdot (\alpha + \beta\zeta^{-1} + \gamma\zeta^{-2})\xi^{-1} + C_1(\alpha)(\alpha + \beta\zeta^{-1})\xi^{-2} \\ & \quad + C_2(\alpha, \beta)\xi^{-3} + C_3(\alpha, \beta, \gamma)\zeta^{-2}\xi^{-2} + O(\xi^{-4}\log\xi), \end{aligned} \tag{3.23}$$

where

$$a_j = -\int_0^\infty a'(\tau)\tau^j d\tau > 0, \quad j = 1, 2. \tag{3.24}$$

Consider the case where $v_+ < v_-$. From (3.2),

$$C_o \equiv -s^2 + \chi'(v_+) = 0. \tag{3.25}$$

We determine the value of α from

$$C_1(\alpha) \equiv \frac{1}{2}\chi''(v_+)\alpha + sa_1\eta'(v_+) = 0. \tag{3.26}$$

Since $\chi''(v_+) < 0$ by (3.3), $\alpha > 0$. Next, we determine the value of $\beta \in \mathcal{R}$ using

$$C_2(\alpha, \beta) \equiv \frac{1}{6}\chi'''(v_+)\alpha^3 + sa_1\eta''(v_+)\alpha^2 - s^2a_2\eta'(v_+)\alpha - sa_1\eta'(v_+)\beta = 0. \tag{3.27}$$

Using (3.26), we see that the coefficient of γ in $C_3(\alpha, \beta, \gamma)$ is given by

$$\chi''(v_+)\alpha + 3sa_1\eta'(v_+) = sa_1\eta'(v_+) > 0. \tag{3.28}$$

Because $C_3(\alpha, \beta, \gamma)$ is linear in γ if we choose $\gamma_+ > \gamma_-$ appropriately, we have

$$C_3(\alpha, \beta, \gamma_\pm) \equiv (\chi''(v_+)\alpha + 3sa_1\eta'(v_+))\gamma_\pm + \frac{1}{2}\chi''(v_+)\beta^2 \gtrless 0. \tag{3.29}$$

Using α, β, and $\gamma_\pm$, we define ϕ_1 and ϕ_2 as follows:

$$\phi_1(\xi) = \begin{cases} v_+ + (\alpha + \beta\zeta^{-1} + \gamma_-\zeta^{-2})\xi^{-1}, & \xi \geq R, \\ \phi_1(R), & \xi < R, \end{cases} \tag{3.30}$$

$$\phi_2(\xi) = \begin{cases} v_+ + (\alpha + \beta\zeta^{-1} + \gamma_+\zeta^{-2})\xi^{-1}, & \xi \geq R, \\ v_-, & \xi < R, \end{cases} \tag{3.31}$$

where R is a large number such that (3.17) holds with the above ϕ_1 and ϕ_2 for $\xi \geq R$. The rest of the proof is the same as in Lemma 3.2.

In the case where $v_+ > v_-$, we see from (3.3) that $\chi''(v_+) > 0$. Therefore, $\alpha < 0$, β, and $\gamma_+ > \gamma_-$ are determined in the same way as before. We define ϕ_1 and ϕ_2 as follows:

$$\phi_1(\xi) = \begin{cases} v_+ + (\alpha + \beta\zeta^{-1} + \gamma_-\zeta^{-2})\xi^{-1}, & \xi \geq R, \\ v_-, & \xi < R, \end{cases} \tag{3.32}$$

$$\phi_2(\xi) = \begin{cases} v_+ + (\alpha + \beta\zeta^{-1} + \gamma_+\zeta^{-2})\xi^{-1}, & \xi \geq R, \\ \phi_2(R), & \xi < R. \end{cases} \tag{3.33}$$

□

Proof of Theorem 3.1. We employ the successive approximations used in [3]. We prove only the case where $s > 0$. Define a sequence of functions $\{V_n\}$, $V_n \in X$, as follows:

$$V_1 = \phi_1, \qquad V_{n+1} = T[V_n] \quad \text{for } n \geq 1, \tag{3.34}$$

where ϕ_1 was defined in Lemmas 3.2 and 3.3. From (3.12) it is easy to see that $\{V_n\}$ is an increasing sequence. Using Lemma 3.1 (ii), we also see $\phi_1 \le V_n \le \phi_2$ for $n \ge 1$, where ϕ_2 is the function defined in Lemmas 3.2 and 3.3. These imply that $\{V_n\}$ is a bounded sequence and for each $\xi \in \mathcal{R}$, $\lim_{n\to\infty} V_n(\xi) = V(\xi)$. Therefore,

$$\phi_1 \le V(\xi) \le \phi_2, \qquad \xi \in \mathcal{R}. \tag{3.35}$$

Now differentiating the relation $\rho(V_{n+1}) = K[V_n], n \ge 1$, with respect to ξ, we have

$$(-s^2 + \sigma'(V_{n+1}))V'_{n+1} = \frac{1}{s}\int_0^\infty a''(\tau)\{\eta(V_n)(\xi + s\tau) - \eta(V_n)(\xi)\}d\tau.$$

This implies

$$|V'_n(\xi)| \le C|\delta|, \qquad \xi \in \mathcal{R} \quad \text{for } n \ge 2, \tag{3.36}$$

where $|\delta| = |v_+ - v_-|$ and C is a positive constant independent of n. So, from Ascoli–Arzelà, $\{V_n\}$ converges uniformly to V on any compact set of $\mathcal{R}$. Therefore, taking the limit as $n \to \infty$ in (3.34), we see V satisfies $V = T[V]$ and, therefore, $V \in X \cap \mathcal{B}^\infty$.

We now show that $V(-\infty) = v_-$. For this purpose set $V(-\infty) = v_*$. Clearly, $v_* \in [v_+, v_-]$. So, suppose $v_* = v_+$. Then, $V(\xi) = v_+$ for $\xi \in \mathcal{R}$. This is a contradiction. Now assume that $v_* \in (v_+, v_-)$. Then, taking the limit as $\xi \to -\infty$ in (2.5) yields

$$-s^2 v_* + \sigma(v_*) = -\int_0^\infty a'(\tau)\eta(v_*)\,d\tau + A = a(0)\eta(v_*) + A.$$

This equation and (2.6) imply

$$s^2 = \frac{\chi(v_*) - \chi(v_-)}{v_* - v_-},$$

which contradicts (1.8). Therefore, $v_* = v_-$ is the only possibility. □

Finally, we show that V satisfies (1.6). From Lemmas 3.2 and 3.3, for large ξ we find that in Case A,

$$V(\xi) = v_+ + \delta e^{-\lambda\xi} + O(e^{-2\lambda\xi}) \tag{3.37}$$

and in Case B,

$$V(\xi) = v_+ + \alpha\xi^{-1} + \beta\xi^{-2}\log\xi + O(\xi^{-3}\log^2\xi). \tag{3.38}$$

Therefore, $w_+(\tau) = e^{-\lambda s\tau}$ in Case A and $w_+(\tau) = 1$ in Case B. This completes the proof of Theorem 3.1. □

4. Uniqueness. In this section we shall prove the uniqueness of the shock profiles for (1.1). We discuss the case where $s > 0$. In Case A we have the following theorem.

THEOREM 4.1 ([3]). *Suppose that Assumption* 1.1 *holds. Let* $s > 0$, (1.7), *and* (1.10) *hold. Consider Case* A. *Suppose there exist two smooth and strictly monotone solutions* $V_1(\xi)$ *and* $V_2(\xi)$ *for* (2.5). *If*

$$V_1(\xi) - V_2(\xi) = O(e^{-\kappa\xi}) \quad \text{as } \xi \to +\infty \tag{4.1}$$

for a constant $\kappa > \lambda$, then $V_1 \equiv V_2$, where $\lambda > 0$ is determined in Lemma 3.2. *In particular, the solution of the form* (3.37) *constructed in Theorem* 3.1 *is unique.*

In Case B we have the following theorem.

THEOREM 4.2. *Suppose that Assumption* 1.1 *holds. Let $s > 0$,* (1.7), (1.8), *and* (1.10) *hold. We also assume* (3.5) *holds (actually $(1+t)^2 a'(t) \in L^1(0,\infty)$ is enough). Consider Case* B. *Suppose that there exist two smooth and strictly monotone solutions $V_1(\xi)$ and $V_2(\xi)$ for* (2.5) *of the form*

$$V(\xi) = v_+ + \alpha\xi^{-1} + o(\xi^{-1}) \quad \text{as } \xi \to +\infty, \tag{4.2}$$

where α is determined in (3.26). *If*

$$V_1(\xi) - V_2(\xi) = O(\xi^{-m}) \quad \text{as } \xi \to +\infty \tag{4.3}$$

for a constant $m > 2$, then $V_1 \equiv V_2$. In particular, the solution of the form (3.38) *constructed in Theorem* 3.1 *is unique.*

Proof. The difference $V_1 - V_2$ satisfies

$$(V_1 - V_2)(\xi) = -\int_0^\infty a'(\tau)\frac{H(\xi + s\tau)}{-s^2 + S(\xi)}(V_1 - V_2)(\xi + s\tau)d\tau, \tag{4.4}$$

where

$$S(\xi) = \int_0^1 \sigma'(V_1(\xi) + \theta(V_1 - V_2)(\xi))d\theta,$$
$$H(\xi) = \int_0^1 \eta'(V_1(\xi) + \theta(V_1 - V_2)(\xi))d\theta.$$

Set

$$N_k(\xi) = \sup_{\zeta \geq \xi} \zeta^k |(V_1 - V_2)(\zeta)|, \tag{4.5}$$

where $2 < k \leq m$. Estimate (4.4) using $N_k(\xi)$. Then we see

$$\xi^k |(V_1 - V_2)(\xi)| \leq D(\xi)N_k(\xi), \tag{4.6}$$

where

$$D(\xi) = -\int_0^\infty a'(\tau)\frac{H(\xi + s\tau)}{-s^2 + S(\xi)}(1 + s\tau\xi^{-1})^{-k}d\tau. \tag{4.7}$$

Consider the asymptotic behavior of $D(\xi)$ as $\xi \to +\infty$. Since V_1 and V_2 satisfy (4.2), for $\forall\theta \in [0,1]$,

$$V_1(\xi) + \theta(V_1 - V_2)(\xi) = v_+ + \alpha\xi^{-1} + o(\xi^{-1}) \quad \text{as } \xi \to +\infty.$$

So,

$$-s^2 + S(\xi) = -s^2 + \sigma'(v_+) + \sigma''(v_+)\alpha\xi^{-1} + o(\xi^{-1}) \quad \text{as } \xi \to +\infty.$$

This implies

$$\frac{1}{-s^2 + S(\xi)} = \frac{1}{-s^2 + \sigma'(v_+)} - \frac{\sigma''(v_+)\alpha}{(-s^2 + \sigma'(v_+))^2}\xi^{-1} + o(\xi^{-1}) \quad \text{as } \xi \to +\infty.$$

Similarly, as $\xi \to +\infty$,

$$H(\xi + s\tau)(1 + s\tau\xi^{-1})^{-k}$$
$$= \eta'(v_+) + (\eta''(v_+)\alpha - ks\tau\eta'(v_+))\xi^{-1} + o(\xi^{-1}) + O(\tau^2\xi^{-2}).$$

We note that o and O are uniform in τ large. From the above computation, we obtain as $\xi \to +\infty$,

$$\frac{H(\xi + s\tau)}{-s^2 + S(\xi)}(1 + s\tau\xi^{-1})^{-k}$$
$$= \frac{\eta'(v_+)}{-s^2 + \sigma'(v_+)}$$
$$- \frac{1}{-s^2 + \sigma'(v_+)} \left\{ \left(\frac{\eta'(v_+)}{-s^2 + \sigma'(v_+)}\sigma''(v_+) - \eta''(v_+) \right) \alpha + ks\tau\eta'(v_+) \right\} \xi^{-1}$$
$$+ o(\xi^{-1}) + O(\tau^2\xi^{-2}).$$

Substitute this in (4.7) and evaluate the integral. In doing so, observe the following relations:

$$-s^2 + \sigma'(v_+) = a(0)\eta'(v_+),$$

$$\frac{1}{-s^2 + \sigma'(v_+)} \left\{ a(0) \left(\frac{\eta'(v_+)}{-s^2 + \sigma'(v_+)}\sigma''(v_+) - \eta''(v_+) \right) \alpha + ksa_1\eta'(v_+) \right\}$$
$$= \frac{\chi''(v_+)\alpha + ksa_1\eta'(v_+)}{a(0)\eta'(v_+)} = \frac{(k-2)sa_1}{a(0)},$$

where (3.26) was used to obtain the second relation. These relations yield

$$D(\xi) = 1 - \frac{(k-2)sa_1}{a(0)}\xi^{-1} + o(\xi^{-1}) \quad \text{as } \xi \to +\infty. \tag{4.8}$$

So, choosing c_o such that $0 < c_o < (k-2)sa_1/a(0)$, we have for large $R > 0$,

$$D(\xi) \le 1 - c_o\xi^{-1} \quad \text{for } \xi \ge R.$$

Choose a value of k such that $2 < k < m$. Since $\xi^k|(V_1 - V_2)(\xi)| \to 0$ as $\xi \to +\infty$, there is a maximum of $\xi^k|(V_1 - V_2)(\xi)|$ on $\xi \ge R$. So, there exists a $\xi_* \ge R$ such that

$$N_k(R) = \xi_*^k|(V_1 - V_2)(\xi_*)|.$$

On the other hand, from (4.6),

$$\xi^k|(V_1 - V_2)(\xi)| \le D(\xi)N_k(R) \quad \text{for } \xi \ge R. \tag{4.9}$$

Therefore, if we set $\xi = \xi_*$, we have $N_k(R) \le D(\xi_*)N_k(R)$. As $D(\xi_*) < 1$, it follows that $N_k(R) = 0$. This implies $V_1(\xi) = V_2(\xi)$ for $\xi \ge R$. Extending into $\xi \le R$ using (4.4), we see $V_1 \equiv V_2$. This completes the proof. □

REFERENCES

[1] J. M. GREENBERG, *Existence of steady shock waves in nonlinear materials with memory*, Arch. Rational Mech. Anal., 24 (1967), pp. 1–21.

[2] ———, *Existence of steady waves for a class of nonlinear dissipative materials*, Quart. Appl. Math., 26 (1968), pp. 27–34.

[3] J. M. GREENBERG AND S. HASTING, *Progressive waves in materials exhibiting long range memory*, SIAM J. Appl. Math., 30 (1976), pp. 31–41.

[4] H. HATTORI AND S. KAWASHIMA, *Nonlinear stability of travelling wave solutions for viscoelastic materials with memory*, J. Differential Equations, to appear.

[5] T. P. LIU, *Nonlinear waves for viscoelasticity with fading memory*, J. Differential Equations, 76 (1988), pp. 26-46.

[6] A. C. PIPKIN, *Shock structure in a viscoelastic fluid*, Quart. Appl. Math., 23 (1966), pp. 297–303.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1143–1156, September 1995

004

STATIONARY SOLUTIONS OF BOUNDARY VALUE PROBLEMS FOR A MAXWELL–BOLTZMANN SYSTEM MODELLING DEGENERATE SEMICONDUCTORS *

A. NOURI† AND F. POUPAUD†

Abstract. One of the most accurate models for carrier transport in semiconductors is based on the Maxwell–Boltzmann system. Degeneracy effects are taken into account by the nonlinearity of the collisions operator. We use two recent techniques developed for the study of kinetic models, upper solutions and mean compactness results, to prove existence of stationary solutions with arbitrary large boundary data, in any kind of geometries.

Key words. boundary value problem, stationary solutions, Vlasov–Maxwell systems, Fermi–Dirac distribution, semiconductor

AMS subject classifications. 35F30, 76P05, 78A35, 82A45

Introduction. For bulk components, the drift diffusion equations give the basic model for the transport of carriers. However, transport phenomena that occur in submicron devices are due to hot and ballistic electrons. In these conditions, it is well known that the drift diffusion model is no longer valid. The physics description needs kinetic models. This paper is devoted to the analysis of one of the most accurate kinetic models, the Maxwell–Boltzmann system.

We use the upper solutions technique of [8] to construct solutions for stationary boundary value problems. In a previous paper [10] we analyzed the Maxwell–Boltzmann system for semiconductors but with a nondegeneracy assumption. Compared to this previous work, the new difficulty is to control the nonlinearity of the collision operator that takes into account the degeneracy effects. The main tools for that are the mean compactness results of [4] and a monotonicity property of the nonlinear collision the operator.

1. The kinetic model and the main result. In kinetic theory, the transport process of charged particles in a self-consistent electromagnetic field is modelled by the Vlasov–Maxwell equations. In semiconductor statistics [2], [3], the distribution function depends on the position x and the wavevector p, instead of the velocity, as in classical theory, in order to take into account some quantum phenomena. Then the velocity and the energy of an electron are given functions of the wavevector. They are related by the relation

$$v(p) = \frac{1}{\hbar}\nabla_p \mathcal{E}(p), \tag{1.1}$$

where $\mathcal{E}(p)$, the energy of the particles, belongs to $(C_b^2(\mathbb{R}^3))^3$, $v(p)$ is the velocity, and $\hbar$ is the reduced Planck constant. For instance, with the parabolic band approximation, we get

$$\mathcal{E}(p) = \hbar^2 \frac{|p|^2}{2m^*}, \qquad v(p) = \frac{\hbar}{m^*}p, \tag{1.2}$$

* Received by the editors March 16, 1993; accepted for publication (in revised form) January 3, 1994.

†Laboratoire J. A. Dieudonné, Unité de Recherche Associée 168 du Centre National de la Recherche Scientifique, Université de Nice Sophia-Antipolis, Parc Valrose, BP 71, F-06108 Nice cédex 2, France.

where m^* is the effective mass of electrons. Then we find the classical identity

$$\mathcal{E} = \frac{1}{2} m^* |v|^2. \tag{1.3}$$

But another model, often used in semiconductor physics, is given implicitly by

$$\mathcal{E}(p) + \alpha \mathcal{E}(p)^2 = \frac{\hbar^2 |p|^2}{2m^*}, \tag{1.4}$$

where α is the coefficient of nonparabolicity. Hence, for the sake of generality, we will consider an arbitrary band diagram $\mathcal{E}(p)$. We only assume that there is a constant β such that for any unitary vector e of $\mathbb{R}^3$, and any positive reals R and γ,

$$\text{meas}\{|p| \leq R \text{ and } |v(p) \cdot e| \leq \gamma\} \leq c(R)\gamma^\beta. \tag{1.5}$$

This means that the velocity v cannot be concentrated along one direction. This assumption is needed for using a compactness property on averages on p of solutions of transport equations. Clearly, it is satisfied if $\mathcal{E}$ is given by (1.2) or (1.4).

Then the distribution $f = f(x,p)$ is determined as follows. Let Ω be an open bounded set of $\mathbb{R}^3$, modelling the device geometry. Let Σ_- denote the subset of the boundary where the velocities are inward:

$$\Sigma_- = \{(x,p) \in \partial\Omega \times \mathbb{R}^3 : v(p) \cdot \nu(x) < 0\}, \tag{1.6}$$

where $\nu(x)$ is the unit outward normal to $\partial\Omega$.

The distribution f solves the following Vlasov–Maxwell equations:

$$v(p) \cdot \nabla_x f(x,p) + F(x,p) \cdot \nabla_p f(x,p) = C(f)(x,p), \quad x \in \Omega, \quad p \in \mathbb{R}^3, \tag{1.7}$$

$$-\Delta_x \phi(x) = \frac{q}{\varepsilon_r}[N(x) - \rho(x)], \quad x \in \Omega, \tag{1.8}$$

$$\nabla_x \wedge B(x) = \mu_r q j(x), \quad \nabla \cdot B(x) = 0, \quad x \in \Omega, \tag{1.9}$$

$$F(x,p) = \frac{q}{\hbar}[\nabla_x \phi(x) - v(p) \wedge B(x)], \quad x \in \Omega, \quad p \in \mathbb{R}^3. \tag{1.10}$$

The constants q, ε_r, and μ_r are, respectively, the charge of the electron, the permittivity, and the permeability of the semiconductor. The function N is the given doping profile. We assume N in $L^\infty(\Omega)$. The operator C is intended to model the collisions between the electrons, impurities, and phonons of the semiconductor [7]. It is defined by

$$C(f)(x,p) = \int_{\mathbb{R}^3} s(p,p')[m(p)f(x,p')(1 - f(x,p)) - m(p')f(x,p)(1 - f(x,p'))]\, dp'. \tag{1.11}$$

The function m is a Maxwellian:

$$m(p) = \exp(-\mathcal{E}(p)/\theta), \tag{1.12}$$

where θ is a physical constant related to the fixed temperature of the semiconductor. The terms $(1 - f)$ in (1.11) come from Pauli's exclusion principle. They express the

fact that there is at most one particle for a given quantum state (x,p). Thus, the physically admissible distribution f will satisfy

$$0 \le f \le 1. \tag{1.13}$$

The function s is given and satisfies

$$s > 0, \qquad s(p,p') = s(p',p), \tag{1.14}$$

$$m(p)s(p,p') \in L^2(\mathbb{R}^6), \tag{1.15}$$

and the collision frequency is assumed to be bounded:

$$\sigma(p) = \int_{\mathbb{R}^3} s(p,p')m(p')\,dp' \in L^\infty(\mathbb{R}^3). \tag{1.16}$$

We recall the null-space of the operator C, which gives the thermal equilibrium distributions (see [9]): it consists of the Fermi–Dirac distributions

$$N(C) = \left\{ n(p) = \left[1 + \exp\left(\frac{\mathcal{E}(p) - \kappa}{\theta}\right)\right]^{-1}; \kappa \in [-\infty, +\infty] \right\}. \tag{1.17}$$

The concentration ρ and the flux j depend on the particle distribution f through the relations

$$\rho(x) = \int_\Omega f(x,p)\,dp, \quad j(x) = \int_\Omega v(p)f(x,p)\,dp, \quad x \in \Omega. \tag{1.18}$$

The system (1.6)–(1.9) is completed with the boundary conditions

$$f(x,p) = g_0(x,p), \qquad (x,p) \in \Sigma_-, \tag{1.19}$$

$$\phi(x) = \phi_0(x), \qquad x \in \partial\Omega, \tag{1.20}$$

$$B(x) \cdot \nu(x) = b(x), \qquad x \in \partial\Omega. \tag{1.21}$$

In order to allow the extension of the boundary data, we assume the following.

(H1) Ω is a smooth bounded set of $\mathbb{R}^3$. Its boundary $\partial\Omega$ is compact and connected.

(H2) $\phi_0 \in H^{1/2}(\partial\Omega) \cap L^\infty(\partial\Omega)$.

(H3) $b \in H^{-1/2}(\partial\Omega)$; $\langle b, 1\rangle_{H^{-1/2},H^{1/2}} = 0$.

Then there are two functions Φ_0 and B_0 such that:

$$\Phi_0 \in H^1(\Omega) \cap L^\infty(\Omega), \quad -\Delta_x\Phi_0 = \frac{q}{\varepsilon_r}N, \quad \Phi_{0/\partial\Omega} = \phi_0,$$

$$B_0 \in H(\text{div, curl, } \Omega), \quad \nabla_x \cdot B_0 = 0, \quad \nabla_x \wedge B_0 = 0, \quad B_0 \cdot \frac{\nu}{\partial\Omega} = b.$$

Finally, we assume that the boundary distribution g_0 is nonnegative and bounded by a Fermi–Dirac distribution:

$$0 \le g_0 \le \left[1 + \exp\left(\frac{\mathcal{E}(p) - \kappa}{\theta}\right)\right]^{-1}. \tag{H4}$$

We now state the main result of this paper.

THEOREM 1.1. *Under the assumptions* (H1)–(H4), *the stationary Vlasov–Maxwell system* (1.7)–(1.10), (1.19)–(1.21) *has at least one solution* (f, ϕ, B) *belonging to* $L^2(\Omega \times \mathbb{R}^3) \times H^1(\Omega) \times H(\text{div, curl, } \Omega)$ *and verifying*

$$0 \le f \le \left[1 + \exp\left(\frac{\mathcal{E}(p) - \frac{q}{\hbar}\Phi_0(x) - \nu}{\theta}\right)\right]^{-1}, \quad \text{where } \nu = \kappa - \frac{q}{\hbar}\|\phi_0\|_\infty. \tag{1.22}$$

Remark. There is no uniqueness of the solution of the system (1.7)–(1.10), (1.19)–(1.21). We give the following counterexample, based on the idea of trapping particles with a potential created by a background charge density. Let n_0 be an arbitrary positive real number. Define ψ by

$$-\Delta_x \psi = \frac{q}{\varepsilon_r} n_0, \qquad \psi/\partial\Omega = 0. \tag{1.23}$$

We let the background charge density N be equal to

$$\int_{\mathbb{R}^3} \left[1 + \exp\left(\frac{\mathcal{E}(p) - \frac{q}{\hbar}\psi(x)}{\theta}\right)\right]^{-1} dp + n_0, \tag{1.24}$$

and we define Φ_1 by

$$-\Delta_x \Phi_1 = \frac{q}{\varepsilon_r} N, \qquad \Phi_{1/\partial\Omega} = 0. \tag{1.25}$$

Then $f_1 = 0$, associated with $\Phi = \Phi_1$, and

$$f_2 = \left[1 + \exp\left(\frac{\mathcal{E}(p) - \frac{q}{\hbar}\psi(x)}{\theta}\right)\right]^{-1},$$

associated with $\Phi_2 = \psi$, are two solutions of (1.7)–(1.10), (1.19)–(1.21).

The paper is organized as follows: in §2, a modified Vlasov–Poisson problem is solved and a maximum principle property is stated, in order to obtain uniform bounds on the concentrations and the fluxes of the modified problem. Then §3 is devoted to the proof of the full stationary Vlasov–Maxwell problem. Finally, §4 details some compactness results used throughout the paper.

2. A modified Vlasov problem. In this section, the electrostatic potential ϕ and the magnetic field B are assumed to be known, such that

$$\phi \in C_b^2(\Omega), \qquad B \in (C_b^1(\Omega))^3. \tag{2.1}$$

Because zero lies in the spectrum of Vlasov operators, uniqueness fails for boundary value problems. Therefore [1], we add an absorption term and solve the following system:

$$\alpha f(x,p) + v(p) \cdot \nabla_x f(x,p) + F(x,p) \cdot \nabla_p f(x,p) = C(f)(x,p), \quad x \in \Omega, \quad p \in \mathbb{R}^3;$$

$$f(x,p) = g_0(x,p), \qquad (x,p) \in \Sigma_-, \tag{2.2}$$

where F is given by (1.10).

THEOREM 2.1. *Under the assumption* (H4) *the problem* (2.2) *has, for every* $\alpha > 0$, *a unique solution in the set of the square integrable functions on* $\Omega \times \mathbb{R}^3$ *that satisfies*

$$0 \le f \le G_{\phi,\nu}.$$

Let us introduce the Maxwell–Boltzmann distribution

$$G_{\phi,\nu}(x,p) = \left[1 + \exp\left(\frac{\mathcal{E}(p) + \frac{q}{\hbar}\phi(x) - \nu}{\theta}\right)\right]^{-1}, \qquad \nu \in [-\infty, +\infty]. \tag{2.3}$$

This distribution solves the Vlasov equation and will be used as an upper solution in the proof of the existence of a solution of (2.2). It will provide a priori estimates on the density ρ and the flux j that will be useful in the following. In order to obtain some maximum principle, we assumed (H4); it is then possible to bound g_0 by $G_{\phi,\nu}$, where

$$\nu = \kappa - \frac{q}{\hbar}\|\phi_0\|_\infty. \tag{2.4}$$

Next we prove the uniqueness of the solution of (2.2), using a monotonicity property of the collision operator. A similar strategy of proof has been used in [8].

Let us prove the existence of a solution of (2.2). We need the following compactness result, which is an easy variant of the mean compactness results of [4], [5], and [6].

PROPOSITION 2.2. *Let* $(f_n), (g_n)$, *and* (h_n) *be bounded sequences of* $L^2(\Omega \times \mathbb{R}^3)$ *that satisfy:*

$$v(p) \cdot \nabla_x f_n = \operatorname{div}_p g_n + h_n \tag{2.5}$$

in the sense of distributions.

Then, for any Hilbert–Schmidt operator K defined on $L^2(\mathbb{R}^6)$, the sequence $(K(f_n(x,\cdot)))$ is relatively compact in $L^2(\Omega \times \mathbb{R}^3)$.

The proof of Proposition 2.2 is given in §4.

Proof of Theorem 2.1. The system (2.2) to be solved gives

$$[\alpha + \sigma(p) + \lambda(f)]f(x,p) + v(p)\cdot\nabla_x f(x,p) + F(x,p)\cdot\nabla_p f(x,p) = \mu(f)(x,p) \quad \text{on } \Omega\times\mathbb{R}^3,$$

$$f/\Sigma_- = g_0, \tag{2.6}$$

with σ defined as in (1.16) and

$$\lambda(f)(x,p) = \int_{\mathbb{R}^3} s(p,p')[m(p) - m(p')]f(x,p')\,dp', \tag{2.7}$$

$$\mu(f)(x,p) = \int_{\mathbb{R}^3} s(p,p')m(p)f(x,p')\,dp'. \tag{2.8}$$

Let us denote

$$\gamma(f)(x,p) = (\lambda - \mu)(f)(x,p) + \sigma(p) = \int_{\mathbb{R}^3} s(p,p')m(p')[1 - f(x,p')]\,dp'. \tag{2.9}$$

Solving (2.6) results in determining a fixed point of the following operator T. Let us denote

$$X = \{(L,M) \in L^2(\Omega\times\mathbb{R}^3)\times L^2(\Omega\times\mathbb{R}^3) : 0 \le M \le \mu(G_{\phi,\nu}) \text{ and } L-M+\sigma \ge \gamma(G_{\phi,\nu})\}.$$

Clearly, X is a closed convex set of $L^2(\Omega\times\mathbb{R}^3)\times L^2(\Omega\times\mathbb{R}^3)$. For every (L,M) in X, let f be the unique solution in $L^2(\Omega\times\mathbb{R}^3)$ of

$$[\alpha+\sigma(p)+L(x,p)]f(x,p)+v(p)\cdot\nabla_x f(x,p)+F(x,p)\cdot\nabla_p f(x,p) = M(x,p) \quad \text{on } \Omega\times\mathbb{R}^3,$$

$$f/\Sigma_- = g_0. \tag{2.10}$$

f is well defined since $\alpha+\sigma+L$ is positive and v and F belong to $C^1(\mathbb{R}^3)^3$ and $C^1(\Omega\times\mathbb{R}^3)^3$, respectively. Moreover, since M and g_0 are nonnegative, then $f \ge 0$.

We define the operator T on X by

$$T(L,M) = (\lambda(f), \mu(f)).$$

Let us now show that T maps X in X. $G_{\phi,\nu} - f$ is a solution of

$$(\alpha+\sigma+L)(G_{\phi,\nu}-f)+v\cdot\nabla_x(G_{\phi,\nu}-f)+F\cdot\nabla_p(G_{\phi,\nu}-f) = \alpha G_{\phi,\nu}+LG_{\phi,\nu}-M. \tag{2.11}$$

Since, thanks to (1.17),

$$\mu(G_{\phi,\nu}) - (\sigma+\lambda(G_{\phi,\nu}))G_{\phi,\nu} = C(G_{\phi,\nu}) = 0, \tag{2.12}$$

we obtain

$$\alpha G_{\phi,\nu}+LG_{\phi,\nu}-M = \alpha G_{\phi,\nu}+[\sigma+L-M-\gamma(G_{\phi,\nu})]G_{\phi,\nu}+(\mu(G_{\phi,\nu})-M)(1-G_{\phi,\nu}) \ge 0.$$

Moreover, the boundary condition on $G_{\phi,\nu} - f$ gives us

$$(G_{\phi,\nu} - f)/\Sigma_- = G_{\phi,\nu} - g_0, \tag{2.13}$$

which, from hypothesis (H4), is nonnegative. Hence $G_{\phi,\nu} - f \ge 0$, so we have

$$0 \le f \le G_{\phi,\nu} \quad \text{on } \Omega\times\mathbb{R}^3. \tag{2.14}$$

Since μ and γ are, respectively, increasing and decreasing,

$$0 \le \mu(f) \le \mu(G_{\phi,\nu}) \tag{2.15}$$

and

$$\gamma(f) = \sigma + \lambda(f) - \mu(f) \ge \gamma(G_{\phi,\nu}). \tag{2.16}$$

Finally, $\mu(f)$ belongs to $L^2(\Omega\times\mathbb{R}^3)$ because

$$\|\mu(f)\|_2 \le \|s(p,p')m(p)\|_{L^2(\mathbb{R}^6)}\|f\|_2 \le c\|f\|_2. \tag{2.17}$$

A similar proof establishes that $\lambda(f)$ belongs to $L^2(\Omega\times\mathbb{R}^3)$.

Let us prove the continuity of the operator T. Let $(L_n), (M_n)$ be convergent sequences in $L^2(\Omega \times \mathbb{R}^3)$ towards L and M, respectively. Knowing that g_0 belongs to $L^2(\Sigma_-; -v \cdot \nu(x) dv d\sigma(x))$, the associated solutions f_n of the system (2.10) form a bounded sequence of $L^2(\Omega \times \mathbb{R}^3)$; therefore, there is a subsequence (f_{n_k}) of (f_n) which converges to some f in $L^2(\Omega \times \mathbb{R}^3)$ weak. Passing to the limit in (2.10) as n_k tends to infinity gives

$$[\alpha + \sigma(p) + L(x,p)] f(x,p) + v(p) \cdot \nabla_x f(x,p) + F(x,p) \cdot \nabla_p f(x,p) = M(x,p) \quad \text{on } \Omega \times \mathbb{R}^3,$$

$$f/\Sigma_- = g_0, \tag{2.18}$$

thus f is unique and the complete sequence (f_n) converges weakly to f in $L^2(\Omega \times \mathbb{R}^3)$.

Moreover, λ and μ being linear functionals, $(\lambda(f_n))$ and $(\mu(f_n))$, respectively, converge weakly to $\lambda(f)$ and $\mu(f)$ in $L^2(\Omega \times \mathbb{R}^3)$.

In view of (2.14), f_n satisfies

$$0 \le f_n \le G_{\phi,\nu}. \tag{2.19}$$

Thus (f_n) is bounded in $L^\infty(\Omega \times \mathbb{R}^3)$. On the other hand,

$$v(p) \cdot \nabla_x f_n(x,p) = -\nabla_p [F(x,p) f_n(x,p)] + g_n, \tag{2.20}$$

with $(f_n), (F f_n)$, and $(g_n) = (-(\alpha + \sigma - L_n) f_n + M_n)$ bounded in $L^2(\Omega \times \mathbb{R}^3)$.

Then, according to Proposition 2.2, $(\lambda(f_n))$ and $(\mu(f_n))$ belong to a compact set of $L^2(\Omega \times \mathbb{R}^3)$. So, in view of the weak convergence of $(\lambda(f_n))$ and $(\mu(f_n))$ in $L^2(\Omega \times \mathbb{R}^3)$, the complete sequence $(\lambda(f_n))$ and $(\mu(f_n))$, respectively, converges to $\lambda(f)$ and $\mu(f)$ in $L^2(\Omega \times \mathbb{R}^3)$. This proves the continuity of T.

Let us show the compactness of T. If (L_n) and (M_n) are bounded in $L^2(\Omega \times \mathbb{R}^3)$, the associated sequence (f_n) is bounded in $L^2(\Omega \times \mathbb{R}^3)$, so Proposition 2.2 implies that $T(L_n, M_n)$ belongs to a compact set of $L^2(\Omega \times \mathbb{R}^3)$. Therefore, the Schauder fixed point theorem gives the existence of a solution of (2.2).

We now prove the uniqueness of the solution of (2.2). Let f and g be two solutions of (2.2). A small computation leads to

(2.21)

$$C(f) - C(g) = -\int_{\mathbb{R}^3} s(p,p') \{(f-g)[m'(1-f') + mg'] - (f'-g')[m(1-f) + m'g]\} \, dp',$$

where f' denotes $f(x,p')$.

For any function h and with the help of the symmetry of s, we get

$$\int_{\mathbb{R}^3} [C(f) - C(g)] h \, dp = -\int_{\mathbb{R}^3} \int_{\mathbb{R}^3} s(h-h')(f-g)[m'(1-f') + mg'] \, dp \, dp'. \tag{2.22}$$

For any small and positive δ, we define the odd function sg^δ and the function abs^δ by

$$sg^\delta(x) = \begin{cases} \frac{4}{\delta}\left(x - \frac{1}{\delta}x^2\right) & \text{if } x \in \left[0, \frac{\delta}{2}\right], \\ 1 & \text{if } x \ge \frac{\delta}{2}, \end{cases} \tag{2.23}$$

and $abs^\delta(x) = x sg^\delta(x)$.

We choose $h = sg^\delta(f-g)$. In view of (2.14), we obtain

$$(h-h')(f-g) = abs^\delta(f-g) - sg^\delta(f'-g')(f-g) \geq -2\delta. \tag{2.24}$$

Then (2.22) implies

$$\int_{\mathbb{R}^3} [C(f) - C(g)] sg^\delta(f-g)\, dp \leq 2c\delta, \tag{2.25}$$

where c is the constant given by

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} s[m'(1-f') + mg']\, dp\, dp' \leq \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} s[m'+m]\, dp\, dp' = c. \tag{2.26}$$

Since f and g are solutions of (2.2),

$$\alpha(f-g) + v(p)\cdot\nabla_x(f-g) + F\cdot\nabla_p(f-g) = C(f) - C(g).$$

Multiplying this equation by $sg^\delta(f-g)$ and integrating it on $\Omega\times\mathbb{R}^3$ leads to

$$\alpha\int_{\Omega\times\mathbb{R}^3}(f-g)sg^\delta(f-g) + \int_\Sigma v(p)\cdot\nu\, abs^\delta(f-g) = \int_{\Omega\times\mathbb{R}^3}[C(f)-C(g)]sg^\delta(f-g). \tag{2.27}$$

But

$$f - g = 0 \quad \text{on } \Sigma_-.$$

Then

$$\alpha\int_{\Omega\times\mathbb{R}^3}(f-g)sg^\delta(f-g) \leq c\delta. \tag{2.28}$$

As δ tends to zero, we get

$$\alpha\int_{\Omega\times\mathbb{R}^3}|f-g| = 0, \tag{2.29}$$

so $f = g$.

3. The Vlasov–Maxwell problem. This section is devoted to the proof of Theorem 1.1. First let us give a sketch of this proof. We regularize the force field and add an absorption term in order to be in the frame of §2. We solve the regularized problem by means of the Schauder fixed point theorem and obtain a solution $(f_\alpha, \phi_\alpha, B_\alpha)$. Seeing that the potentials ϕ_α and Φ_0 satisfy

$$-\Delta_x\phi_\alpha = \frac{q}{\varepsilon_r}(N-\rho_\alpha), \quad \Delta_x\Phi_0 = \frac{q}{\varepsilon_r}N, \quad \Phi_{0/\partial\Omega} = \phi_0 = \phi_{\alpha/\partial\Omega},$$

the following inequality holds:

$$q\phi_\alpha \leq q\Phi_0.$$

It follows that the maximum principle of §2 applies to (f_α, ϕ_α): uniform bounds on the flux j_α and the concentration ρ_α are obtained, which gives compactness properties for F in $L^2(\Omega\times\mathbb{R}^3)$.

Finally, we pass to the limit in the modified system: we overcome the problem of passing to the limit in the nonlinear collision operator with the help of the compactness result given in Proposition 2.2.

We define a regularized force field in the following way. For any $\alpha > 0$,

$$F_\alpha = F_\alpha(\phi, B) = \frac{q}{\hbar}[\nabla_x \psi_\alpha(\phi) - v(p) \wedge H_\alpha(B)], \tag{3.1}$$

where the modified magnetic field H_α is obtained by regularizing in the classical way:

$$H_\alpha = \bar{B} * \zeta_\alpha.$$

$\bar{B}$ is the prolongation of B by zero outside Ω, and ζ_α is a regularizing sequence:

$$\zeta_\alpha(x) = \frac{1}{\alpha^3}\,\zeta\Big(\frac{x}{\alpha}\Big), \quad \int_{\mathbb{R}^3} \zeta(x)\,dx = 1, \quad \zeta \in C_0^\infty(\mathbb{R}^3).$$

To get a regularized potential ψ_α of ϕ such that ψ_α belongs to $C_b^2(\Omega)$, and $\psi_{\alpha/\partial\Omega} = \phi_0$ and $q\phi_\alpha \le q\Phi_0$ as soon as $\phi/\partial\Omega = \phi_0$ and $q\phi \le q\Phi_0$, we choose

$$\psi_\alpha = \Phi_0 + (I - \alpha\Delta)^{-2}(\phi_0 - \Phi_0),$$

where the operator Δ is considered an unbounded operator on $L^2(\Omega)$ whose domain is $H^2(\Omega) \cap H_0^1(\Omega)$. We remark that $(I - \alpha\Delta)^{-2}(\phi_0 - \Phi_0)$ belongs to $H^4(\Omega)$, then to $C_b^2(\Omega)$. Thus, in order that ψ_α belongs to $C_b^2(\Omega)$, we assume that

(H5) $\Phi_0 \in C_b^2(\Omega)$.

Then the properties of the above regularization are summarized in the following lemma.

LEMMA 3.1. *The map $F_\alpha = F_\alpha(\phi, B)$ is continuous from $H^1(\Omega) \times L^2(\Omega)$ into $C_b^1(\Omega \times \mathbb{R}^3)$. For any potential ϕ such that*

$$\phi \in H^1(\Omega), \quad \phi/\partial\Omega = \phi_0, \quad \text{and} \quad q\phi \le q\Phi_0,$$

the modified potential $\psi_\alpha = \psi_\alpha(\phi)$ satisfies

$$\psi_\alpha \in C_b^2(\Omega), \quad \psi_{\alpha/\partial\Omega} = \phi_0, \quad q\phi_\alpha \le q\Phi_0.$$

Furthermore, for any sequence (α_n, ϕ_n, B_n) such that

$$\alpha_n \to 0,$$
$$(\phi_n) \text{ is uniformly bounded in } H^2(\Omega), \quad \phi_{n/\partial\Omega} = \phi_0, \quad \phi_n \to \phi \quad \text{in } H^1(\Omega),$$
$$B_n \to B \quad \text{in } L^2(\Omega),$$

the regularized force field $F_{\alpha_n}(\phi_n, B_n)$ converges towards $F = \frac{q}{\hbar}(\nabla_x\phi - v(p) \wedge B)$ in $L^2_{\text{loc}}(\bar{\Omega} \times \mathbb{R}^3)$.

For a proof of this lemma, we refer the reader to [8].

The modified Vlasov–Maxwell system. The regularized problem is

$$\begin{aligned} &\alpha f_\alpha + v(p) \cdot \nabla_x f_\alpha + F_\alpha \cdot \nabla_p f_\alpha = C(f_\alpha) \quad \text{on } \Omega \times \mathbb{R}^3, \\ &f_{\alpha/\Sigma_-} = g_0. \end{aligned} \tag{3.2}$$

The regularized force field is given by (3.1). The potential solves

$$
\begin{aligned}
&-\Delta_x \phi_\alpha = \frac{q}{\varepsilon_r}(N - \rho_\alpha),\\
&\rho_\alpha(x) = \int_{\mathbb{R}^3} f_\alpha(x,p)\,d\rho,\\
&\phi_\alpha(x) = \phi_0(x) \quad \text{on } \partial\Omega.
\end{aligned}
\tag{3.3}
$$

The magnetostatic problem has to be modified because the flux of a solution of (3.2) is no longer divergence free. Instead we obtain

$$\alpha\rho_\alpha + \nabla_x \cdot j_\alpha = 0.$$

But (3.3) shows that

$$\nabla_x \cdot \left[j_\alpha + \alpha \frac{\varepsilon_r}{q} \nabla_x (\phi_\alpha - \Phi_0) \right] = 0.$$

Therefore, we define the new magnetostatic problem by

$$
\begin{aligned}
&\nabla_x \wedge B_\alpha = \mu_r q \left[j_\alpha + \alpha \frac{\varepsilon_r}{q} \nabla_x (\phi_\alpha - \Phi_0) \right],\\
&\nabla_x \cdot B_\alpha = 0,\\
&j_\alpha(x) = \int_{\mathbb{R}^3} v(p) f_\alpha(x,p)\,dp,\\
&B_\alpha(x) \cdot \nu(x) = b(x) \quad \text{on } \partial\Omega.
\end{aligned}
\tag{3.4}
$$

PROPOSITION 3.2 (*existence for the modified problem*). *Let $\alpha > 0$. Under the hypotheses* (H1)–(H5), *the modified Vlasov–Maxwell system* (3.2)–(3.4) *has at least one solution $(f_\alpha, \phi_\alpha, B_\alpha) \in L^2(\Omega \times \mathbb{R}^3) \times H^2(\Omega) \times H^1(\Omega)$, which satisfies uniformly with respect to α:*

$$0 \le f_\alpha \le \left[1 + \exp\left(\frac{\mathcal{E}(p) - \frac{q}{\hbar}\Phi_0(x) - \nu}{\theta} \right) \right]^{-1}. \tag{3.5}$$

ϕ_α is uniformly bounded in $H^2(\Omega)$; $B_\alpha - B_0$ is uniformly bounded in $H^1(\Omega)$.

Proof of Proposition 3.2. Let Ξ be the following nonempty convex closed set of $H^1(\Omega) \times L^2(\Omega)$:

$$\Xi = \{(\phi, B) \in H^1(\Omega) \times L^2(\Omega) : \phi/\partial\Omega = \phi_0 \text{ and } q\phi \le q\Phi_0\}.$$

We define a map on Ξ in the following way. For every (ϕ, B) in Ξ, let $f_{\phi,B}$ be the unique solution of the modified Vlasov–Maxwell problem (3.2). Then we define

$$
\begin{aligned}
\rho_{\phi,B}(x) &= \int_{\mathbb{R}^3} f_{\phi,B}(x,p)\,dp,\\
j_{\phi,B}(x) &= \int_{\mathbb{R}^3} v(p) f_{\phi,B}(x,p)\,dp.
\end{aligned}
$$

Let (ϕ_1, B_1) be the solution of (3.3), (3.4) with the corresponding concentration and flux. The map Γ is defined by $\Gamma(\phi, B) = (\phi_1, B_1)$. The following property and the

Schauder fixed point theorem establish the existence of a solution $(f_\alpha, \phi_\alpha, B_\alpha)$ of (3.2)–(3.4).

Property 3.3. $\Gamma : \Xi \to \Xi$ is continuous and compact for the topology of $H^1(\Omega \times \mathbb{R}^3) \times L^2(\Omega \times \mathbb{R}^3)$.

Let us first show the compactness of Γ. From (H2) and the maximum principle established in Theorem 2.1, we know that

$$
\begin{aligned}
0 \le f_{\phi,B} &\le \left[1 + \exp\left(\frac{\mathcal{E}(p) - \frac{q}{\hbar}\phi(x) - \nu}{\theta}\right)\right]^{-1} \\
&\le \left[1 + \exp\left(\frac{\mathcal{E}(p) - \frac{q}{\hbar}\Phi_0(x) - \nu}{\theta}\right)\right]^{-1}.
\end{aligned} \tag{3.6}
$$

Therefore, $\rho_{\phi,B}$ and $j_{\phi,B}$ are uniformly bounded by a constant c depending only on $\|\Phi_0\|_\infty$ and ν:

$$0 \le \rho_{\phi,B} \le c \quad \text{and} \quad |j_{\phi,B}| \le c.$$

Then the solution η in $H_0^1(\Omega)$ of

$$-\Delta_x \eta = -\frac{q}{\varepsilon_r}\rho_{\phi,B}$$

is uniformly bounded in $H^2(\Omega)$ and satisfies $q\eta \le 0$. Hence the function $\phi_1 = \Phi_0 + \eta$ lies in a bounded set of $H^2(\Omega)$, which is a compact set of $H^1(\Omega)$ and satisfies

$$\phi_{1/\partial\Omega} = \phi_0, \qquad q\phi_1 \le q\Phi_0.$$

The function

$$j_{\phi,B} + \alpha\frac{\varepsilon_r}{q}\nabla_x(\phi_1 - \Phi_0)$$

belongs to a bounded set of $L^2(\Omega)$ and satisfies

$$\nabla_x \cdot \left[j_{\phi,B} + \alpha\frac{\varepsilon_r}{q}\nabla_x(\phi_1 - \Phi_0)\right] = 0.$$

Then the solution D of

$$
\begin{aligned}
&\nabla_x \wedge D = \mu_r q\left[j_{\phi,B} + \alpha\frac{\varepsilon_r}{q}\nabla_x(\phi_1 - \Phi_0)\right], \qquad \nabla_x \cdot D = 0, \\
&D \cdot \nu = 0 \quad \text{on } \partial\Omega
\end{aligned}
$$

belongs to a bounded set of $H^1(\Omega)$. Thus $B_1 = B_0 + D$ belongs to a compact set of $L^2(\Omega)$ and we have proved that (ϕ_1, B_1) lies in a compact subset of Ξ.

Let us show the continuity of Γ. Let (ϕ_n, B_n) in Ξ be such that ϕ_n converges to ϕ in $H^1(\Omega)$ and B_n converges to B in $L^2(\Omega)$. Then $F_\alpha(\phi_n, B_n)$ converges towards $F_\alpha(\phi, B)$ in $C_b^1(\Omega \times \mathbb{R}^3)$. (3.6) implies that there is a subsequence $f_n = f_{\phi_n,B_n}$ which converges weakly in $L^2(\Omega \times \mathbb{R}^3)$ towards some f. Then, with the help of Proposition 2.2, $\lambda(f_n)$ and $\mu(f_n)$ converge towards $\lambda(f)$ and $\mu(f)$, respectively, in $L^2(\Omega \times \mathbb{R}^3)$, so $C(f_n)$ converges to $C(f)$ in the distributional sense. It follows that f solves the Poisson equation associated with (ϕ, B) and hence is equal to $f_{\phi,B}$. Then ρ_n and j_n, respectively, converge weakly to $\rho_{\phi,B}$ and $j_{\phi,B}$ in $L^2(\Omega)$. Since the sequence $\Gamma(\phi_n, B_n) = (\phi_{n,1}, B_{n,1})$

belongs to a compact set of $H^1(\Omega)\times L^2(\Omega)$, the last convergences show that $(\phi_{n,1}, B_{n,1})$ converges to (ϕ_1, B_1) in $H^1(\Omega)\times L^2(\Omega)$.

Therefore, the Schauder fixed point theorem applies, which shows the existence of a solution $(f_\alpha, \phi_\alpha, B_\alpha)$ of (3.2)–(3.4). Moreover, f_α satisfies

$$0 \le f_\alpha \le \left[1+\exp\left(\frac{\mathcal{E}(p)-\frac{q}{\hbar}\Phi_0(x)-\nu}{\theta}\right)\right]^{-1},$$

as $f_{\phi,B}$ does.

We then deduce from the proof of the compactness of Γ that ϕ_α belongs to a bounded set of $H^2(\Omega)$ and that $B_\alpha - B_0$ belongs to a bounded set of $H^1(\Omega)$.

We now prove the existence of a solution of the complete Vlasov–Maxwell system. We first assume that Φ_0 satisfies (H5). Let $(f_\alpha, \phi_\alpha, B_\alpha)$ be the solution of the modified problem for any $\alpha > 0$. In view of the uniform estimates (3.5), there is a subsequence α_n converging to 0, such that $(f_{\alpha_n}, \phi_{\alpha_n}, B_{\alpha_n})$, denoted by (f_n, ϕ_n, B_n), satisfies

$$f_n \to f \text{ weakly} \quad \text{in } L^2(\Omega\times\mathbb{R}^3),$$
$$(\phi_n) \text{ is uniformly bounded in } H^2(\Omega), \quad \phi_{n/\partial\Omega} = \phi_0, \quad \phi_n \to \phi \quad \text{in } H^1(\Omega),$$
$$B_n \to B \quad \text{in } L^2(\Omega).$$

Then

$$F_{\alpha_n}(\phi_n, B_n) \to F = \frac{q}{\hbar}(\nabla_x\phi - v\wedge B) \quad \text{in } L^2_{\text{loc}}(\bar\Omega\times\mathbb{R}^3).$$

As in the proof of the continuity of Γ, $C(f_n)$ converges to $C(f)$ in the distributional sense. Hence f is a solution of (1.7)–(1.10). Moreover, in view of (3.5) and the choice of the constant ν,

$$\rho_n(x) = \int_{\mathbb{R}^3} f_n(x,p)\,dp \quad \text{and} \quad j_n(x) = \int_{\mathbb{R}^3} v(p) f_n(x,p)\,dp$$

are uniformly bounded, so that

$$\rho_n(x) \to \rho(x) = \int_{\mathbb{R}^3} f(x,p)\,dp \quad \text{in } L^\infty(\Omega) \text{ weak star},$$

and

$$j_n(x) \to j(x) = \int_{\mathbb{R}^3} v(p) f(x,p)\,dp \quad \text{in } L^\infty(\Omega) \text{ weak star}.$$

Then it is straightforward to pass to the limit in (3.3), (3.4) and obtain a solution of the Vlasov–Maxwell system. To get rid of the restriction (H5), we introduce a sequence $\Phi_{0,n}$ such that

$$\Phi_{0,n} \in C_b^2(\Omega), \quad \Phi_{0,n} \to \Phi_0 \quad \text{in } H^1(\Omega), \quad \|\Phi_0\|_\infty \le c_1$$

and pass to the limit of the corresponding solutions (f_n, ϕ_n, B_n).

Appendix: A compactness result. This section is devoted to the proof of Proposition 2.2, which we now restate.

PROPOSITION 2.2. *Let $(f_n), (g_n)$, and (h_n) be bounded sequences of $L^2(\Omega\times\mathbb{R}^3)$ that satisfy*

$$v(p)\cdot\nabla_x f_n = \operatorname{div}_p g_n + h_n \tag{4.1}$$

in the sense of distributions. Then for any Hilbert–Schmidt operator K defined on $L^2(\mathbb{R}^3)$, the sequence $(K(f_n(x,\cdot))$ is relatively compact in $L^2(\Omega \times \mathbb{R}^3)$.

Proof of Proposition 2.2. Since K is a Hilbert–Schmidt operator, there is a kernel ϕ in $L^2(\mathbb{R}^6)$ such that

$$K : f \to K(f)(x,p) = \int_{\mathbb{R}^3} \phi(p,q) f(x,q)\, dq. \tag{4.2}$$

Let ϕ^N be a sequence converging to ϕ in $L^2(\mathbb{R}^6)$ and verifying

$$\phi^N(p,p') = \sum_{i=1}^{i=M} \psi_i^N(p)\varphi_i^N(p'), \tag{4.3}$$

where ψ_i^N and φ_i^N are compactly supported and indefinitely differentiable. The Hilbert–Schmidt operator K is the uniform limit of the integral operators K^N of kernel ϕ^N, since the norm of $K^N - K$ is the norm of $\phi^N - \phi$ in $L^2(\mathbb{R}^6)$. From classical compactness results (see [4], [5], and [6]), the sequences

$$(\psi_i^N(p) \int_{\mathbb{R}^3} f_k(x,p')\psi_i^N(p')\, dp')_{k\geq 1}$$

belong to a compact set of $L^2(\mathbb{R}^6)$, so finite sums of such sequences also belong to a compact set of $L^2(\mathbb{R}^6)$. By the diagonal process we construct a subsequence (f_{k_p}) such that the sequence $(K^N(f_{k_p}))_{p\geq 1}$ converges.

Let us show that $(K(f_{k_p}))_{p\geq 1}$ is a Cauchy sequence in $L^2(\mathbb{R}^6)$:

$$\begin{aligned} \|K(f_{k_p}) - K(f_{k_q})\| &\leq \|(K^N - K)(f_{k_p} - f_{k_q})\| + \|K^N(f_{k_p}) - K^N(f_{k_q})\| \\ &\leq 2\|K^N - K\|M + \|K^N(f_{k_p}) - K^N(f_{k_q})\|, \end{aligned} \tag{4.4}$$

where M is a bound of $\|f_k\|_{L^2(\mathbb{R}^6)}$. $\varepsilon > 0$ being given, there is an integer N_0 such that

$$2\|K^{N_0} - K\|M < \frac{\varepsilon}{2}. \tag{4.5}$$

$(K^{N_0}(f_{k_p}))$ being a Cauchy sequence, there is an integer P such that for every $p \geq P$ and $q \geq P$,

$$\|K^{N_0}(f_{k_p}) - K^{N_0}(f_{k_q})\|_{L^2(\mathbb{R}^6)} \leq \frac{\varepsilon}{2}. \tag{4.6}$$

Using (4.4)–(4.6) leads to

$$\|K(f_{k_p}) - K(f_{k_q})\|_{L^2(\mathbb{R}^6)} \leq \varepsilon, \tag{4.7}$$

which proves that $(K(f_{k_p}))_{p\geq 1}$ is a Cauchy sequence in $L^2(\mathbb{R}^6)$, and ends the proof.

REFERENCES

[1] C. BARDOS, *Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximations; application à l'équation de transport*, Ann. Sci. Ecole Norm. Sup., 4e série, 3 (1970), pp. 185–233.

[2] J. S. BLAKEMORE, *Semiconductor Statistics*, Pergamon Press, Oxford, 1962.

[3] P. DEGOND, F. POUPAUD, B. NICLOT, AND F. GUYOT, *Semiconductor modelling via the Boltzmann equation*, Lectures in Appl. Math., 25 (1990), pp. 51–73.

[4] R. J. DI PERNA AND P. L. LIONS, *Global weak solutions of Vlasov–Maxwell systems*, Comm. Pure Appl. Math., 62 (1989), pp. 729–757.

[5] F. GOLSE, P. L. LIONS, B. PERTHAME, AND R. SENTIS, *Regularity of the moments of the solution of a transport equation*, J. Funct. Anal., 76 (1988), pp. 110–125.

[6] F. GOLSE AND F. POUPAUD, *Limite fluide des équations de Boltzmann des semiconducteurs pour une statistique de Fermi-Dirac*, Asymptotic Anal., 6 (1992), pp. 135–160.

[7] F. POUPAUD, *A half-space problem for a non-linear Boltzmann equation arising in semiconductor statistics*, Math. Models Methods Appl. Sci., 14 (1991), pp. 121–137.

[8] ———, *Boundary value problems for the stationary Vlasov–Maxwell system*, Forum Math., 4 (1992), pp. 499–527.

[9] ———, *On a system of non linear Boltzmann equations of semiconductor physics*, SIAM J. Appl. Math., 50 (1990), pp. 1593–1607.

[10] ———, *Boundary value problems in semiconductors for the stationary Vlasov–Maxwell–Boltzmann equations*, IMA Vol. Math. Appl. 59, Springer-Verlag, Berlin, 1994.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1157–1183, September 1995

005

THE EQUILIBRIUM PLASMA SUBJECT TO SKIN EFFECT*

YONG LIU†

Abstract. An interior free boundary problem of Bernoulli type in an annular region is considered. On the inner (unknown) boundary of the region, the solution of the Laplace equation satisfies the zero Dirichlet condition and a Neumann-type condition. On the outer (given) boundary, the solution assumes a constant value. Existence of solutions is established and uniqueness under some assumption is studied. Examples of nonuniqueness are also given.

Key words. free boundary problem, equilibrium plasma, existence, uniqueness, nonuniqueness

AMS subject classifications. 35R35, 76X05

1. Introduction.

The Tokamak machine is designed to contain and exploit thermonuclear plasma. In the machine, the plasma is confined inside a perfect superconducting shell in which the eddy currents generate the magnetic field necessary for ensuring plasma equilibrium.

A two-dimensional model (the superconducting shell is regarded as an infinite cylinder) was introduced in [13] for the equilibrium plasma subject to a surface current. This simplified model leads to the following free boundary problem.

Given a closed curve Γ (the cross section of the shell) and two positive parameters κ (the magnetic field flux) and ρ_0 (the surface current on the plasma), find a contour γ (the boundary of the plasma region) and a function u (the flux of the magnetic field) such that

$$\triangle u = 0 \qquad \text{in} \quad \Omega_\gamma, \tag{1.1}$$

$$u = \kappa \qquad \text{on} \quad \Gamma, \tag{1.2}$$

$$u = 0 \qquad \text{on} \quad \gamma, \tag{1.3}$$

$$|\nabla u| = \frac{\rho_0}{l_\gamma} \qquad \text{on} \quad \gamma, \tag{1.4}$$

where Ω_γ (the vacuum set) is the region in $\mathbb{R}^2$ bounded by Γ and γ; γ (unknown in advance) lies in the interior of Γ and l_γ is the arc length of γ.

The Tokamak machine is actually toroidal. In this case we need to replace (1.1) and (1.4), respectively, by

$$\triangle u - \frac{1}{x} u_x = 0 \qquad \text{in} \quad \Omega_\gamma \subset R^2 \cap \{x \geq 0\} \tag{1.1'}$$

and

$$\frac{1}{x}|\nabla u| = \frac{\rho_0}{l_\gamma} \qquad \text{on} \quad \gamma. \tag{1.4'}$$

In the special case where Γ is a polygon, existence for (1.1)–(1.4) was proved in [10], [13], and [14]. A related convex-constrained problem was solved in [2]. In [17], an approximated version of (1.1)–(1.4) was studied; this problem coincides with (1.1)–(1.4) in case Γ is a circle.

* Received by the editors May 17, 1993; accepted for publication (in revised form) January 5, 1994.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

The methods used in [10], [13], [14], and [17] rely on conformal mapping, and therefore do not apply to the actual toroidal Tokamak machine.

We are not aware of any uniqueness results regarding (1.1)–(1.4).

In this paper we study the free boundary problem (1.1)–(1.4). We prove that if Γ is a Lipschitz Jordan curve, then there exists a solution pair (γ, u) and γ is an analytic curve. This result can be extended to the problem (1.1′), (1.2), (1.3), and (1.4′) (see Remark 2.2).

Uniqueness for (1.1)–(1.4) is established in this paper under the following additional assumptions:

(1) Γ is symmetric about the coordinate axes;

(2) the portion of Γ lying in the first quadrant is both an x-graph and a y-graph.

Condition (2) is essential for assuring uniqueness. In fact we shall give an example showing nonuniqueness even when Γ undergoes a "small" perturbation from this condition.

The proof of existence is based on a variational approach to the following Bernoulli-type interior free boundary problem.

Let $\Omega \subset \mathbb{R}^2$ be a simply connected bounded domain with Lipschitz boundary $\partial\Omega$. We seek a function $u \geq 0$ such that

$$\triangle u = 0 \quad \text{in} \quad \Omega \cap \{u > 0\}, \tag{1.5}$$

$$u = 1 \quad \text{on} \quad \partial\Omega, \tag{1.6}$$

$$|\nabla u| = \text{const.} \quad \text{on} \quad \Omega \cap \partial\{u > 0\}. \tag{1.7}$$

Alt and Caffarelli [8] studied the variational problem

$$\text{minimize} \quad J_\lambda(v) = \int_\Omega (|\nabla v|^2 + \lambda^2 I_{\{v>0\}})dx, \qquad \lambda > 0 \tag{1.8}$$

over an appropriate class K, where I_D is the characteristic function of the set D. They proved that any minimizer u is Lipschitz continuous in Ω and satisfies (1.5), (1.6) and

$$|\nabla u| = \lambda \quad \text{on} \quad \Gamma_\lambda = \Omega \cap \partial\{u > 0\}. \tag{1.9}$$

They also proved that the free boundary Γ_λ is an analytic curve. This variational formulation was used by Alt, Caffarelli, and Friedman (see [9], [15] and the references therein) to study jet and cavity free boundary problems.

Other variational approaches were introduced later by Acker [3] and Aguilera, Alt, and Caffarelli [7].

Given a constant λ with $0 < \lambda < \text{vol}(\Omega)$, set

$$K_\lambda = \{v \in H^1(\Omega); \quad v \geq 0, v\big|_{\partial\Omega} = 1 \quad \text{and} \quad \text{vol}(\{v > 0\}) = \lambda\}, \tag{1.10}$$

where $\text{vol}(D)$ is the volume of D in the Lebesgue sense, and introduce the Dirichlet integral

$$J(v) = \int_\Omega |\nabla v|^2. \tag{1.11}$$

We seek a function $u \in K_\lambda$ such that

$$J(u) = \min_{v \in K_\lambda} J(v). \tag{1.12}$$

If $\partial\Omega$ is convex, Acker [3] proved that (1.5)–(1.7) is solvable with $\Omega \cap \partial\{u > 0\}$ as a convex curve. This result was generalized later in [7], where it was shown that

any solution u of (1.12) is also Lipschitz in Ω and verifies (1.5) and (1.6), and on $\Gamma_\lambda = \Omega \cap \partial\{u > 0\}$, which is an analytic curve

$$|\nabla u| = c_\lambda > 0 \quad (c_\lambda \text{ is an unprescribed constant}). \tag{1.13}$$

Problem (1.1)–(1.4) appears to be related to both variational problem (1.8) and (1.12). Nevertheless, a special case shows that solutions of (1.1)–(1.4) do not always minimize the functional in (1.8) (see Remark 1.1). We therefore wish to use the variational approach (1.12) and try to "fit" condition (1.4). Since, however, we do not know whether (1.12) has a unique solution, we cannot work directly with the function $\lambda \mapsto c_\lambda$. Instead we shall work with the well-defined function

$$\lambda \longmapsto \min_{v \in K_\lambda} J(v).$$

Several properties of this function are given in §2, and existence for (1.1)–(1.4) is established by exploiting these properties.

In §3 we derive symmetry properties of solutions and use them in §4 to prove uniqueness.

Examples of nonuniqueness are given in §5.

Acker studied the uniqueness of solutions of problem (1.5)–(1.7) in the case where the constant in (1.7) is prescribed (see [4]–[6]). These results have direct application to our problem (1.1)–(1.4). In fact, under some geometric assumption on Γ (not necessarily symmetric), we can prove that (1.1)–(1.4) has a unique solution if ρ_0/κ is small.

In §6, we study an exterior problem related to (1.1)–(1.4) which arises from cryogenics experiments [20]. The free boundary condition is the same as (1.4) but the free boundary γ is in the exterior of Γ. We shall use the variational problem (1.8) to establish existence for any given Lipschitz Jordan curve Γ. Uniqueness and some properties of the solution are also derived.

It is worth mentioning that the corresponding exterior problem of (1.5)–(1.7) has been extensively studied by many authors (see [1], [2], [11], and [23]). In [1], the exterior version of (1.1)–(1.4) was solved when Γ is starlike.

Remark 1.1. In general, solutions of (1.1)–(1.4) do not solve the variational problem (1.8). In fact, in the case where Γ is a unit circle and $\kappa = 1$, a simple computation shows that any minimizer u_λ of (1.8) has the form

$$u_\lambda = \begin{cases} 1 & \text{if} \quad 0 < \lambda < e, \\ \left(1 - \frac{\log r}{\log R_\lambda}\right) I_{\{r > R_\lambda\}} & \text{if} \quad \lambda > e, \end{cases}$$

where R_λ is determined by

$$\frac{1}{|R_\lambda \log R_\lambda|} = \lambda$$

and $R_\lambda \uparrow$ if $\lambda \uparrow$, $R_\lambda \geq e^{-1}$. Since the solution of (1.1)–(1.4) is given by (see Corollary 3.3)

$$u = \left(1 + \frac{\rho_0}{2\pi} \log r\right), \qquad e^{-\frac{2\pi}{\rho_0}} < r < 1,$$

it follows that, if $0 < \rho_0 < 2\pi$, then

$$e^{-\frac{2\pi}{\rho_0}} < e^{-1} \leq R_\lambda \qquad \text{for all} \quad \lambda \geq e$$

and u cannot be a minimizer for (1.8).

2. Existence. Let Γ be a Lipschitz Jordan curve in $\mathbb{R}^2$ which encloses a bounded domain Ω. Introducing the transformation $u \to \frac{1}{\kappa}u$ and setting $I = \rho_0/\kappa$, we can rewrite (1.1)–(1.4) in the form

$$\begin{cases} \triangle u = 0 & \text{in} \quad \Omega_\gamma \subset \mathbb{R}^2, \\ u = 1 & \text{on} \quad \Gamma, \\ u = 0 & \text{on} \quad \gamma, \\ |\nabla u| = \frac{I}{l_\gamma} & \text{on} \quad \gamma. \end{cases} \tag{2.1}$$

THEOREM 2.1. *Given any $I > 0$, (2.1) has a solution (γ, u) with γ an analytic curve contained in Ω.*

We begin by introducing the function

$$m(\lambda) = \min_{v \in K_\lambda} \int_\Omega |\nabla v|^2, \qquad (0 < \lambda < \mathrm{vol}(\Omega)), \tag{2.2}$$

where the set K_λ is given by (1.10). It will be useful to introduce an alternative definition for $m(\lambda)$.

For any compact subset $D \subset\subset \Omega$, set

$$K_D = \{v \in H^1(\Omega); \quad v\big|_{\partial\Omega} = 1, v \equiv 0 \quad \text{on} \quad D\}. \tag{2.3}$$

Let u be the capacitory potential of the region $\Omega \backslash D$:

$$\int_\Omega |\nabla u|^2 = \min_{v \in K_D} \int_\Omega |\nabla v|^2 \equiv \mathrm{Cap}_\Omega D. \tag{2.4}$$

Then

$$m(\lambda) = \min_{\mathrm{vol}(\Omega \backslash D) = \lambda} \mathrm{Cap}_\Omega D. \tag{2.5}$$

Let u_λ be a solution of (1.12). Since $u_\lambda = 1$ on Γ and u_λ is Lipschitz continuous, the set $\{x \in \Omega; u_\lambda > 0\}$ is open and has a component which is connected to Γ. By the maximum principle, there are no other components of this set. Denote by Γ_λ the free boundary $\Omega \cap \partial\{u_\lambda > 0\}$. We shall write c_λ as $|\nabla u_\lambda|(\Gamma_\lambda)$ to emphasize its dependence on u_λ. The following is a key lemma that makes our approach possible.

LEMMA 2.2. $m(\lambda) = |\nabla u_\lambda|(\Gamma_\lambda) l_{\Gamma_\lambda}$.

Proof. By integration by parts, using (1.5) and (1.6), we obtain

$$\int_\Omega |\nabla u_\lambda|^2 = \int_{\{u_\lambda > 0\}} |\nabla u_\lambda|^2 = \int_{\partial\{u_\lambda > 0\}} u_\lambda \frac{\partial u_\lambda}{\partial n} = \int_{\partial\Omega} \frac{\partial u_\lambda}{\partial n}, \tag{2.6}$$

where n is the outward normal. Since

$$0 = \int_{\{u_\lambda > 0\}} \triangle u_\lambda = \int_{\partial\Omega} \frac{\partial u_\lambda}{\partial n} + \int_{\Gamma_\lambda} \frac{\partial u_\lambda}{\partial n} \tag{2.7}$$

and $\partial u_\lambda / \partial n = -|\nabla u_\lambda|$ on Γ_λ,

$$\int_{\partial\Omega} \frac{\partial u_\lambda}{\partial n} = -\int_{\Gamma_\lambda} \frac{\partial u_\lambda}{\partial n} = |\nabla u_\lambda|(\Gamma_\lambda) l_{\Gamma_\lambda} \quad \text{by (1.13).} \tag{2.8}$$

Combining (2.6) and (2.8), the lemma follows.

LEMMA 2.3. *$m(\lambda)$ is continuous for $0 < \lambda < \mathrm{vol}(\Omega)$.*

Proof. We shall show that if $\lambda_j \to \lambda_0$ $(0 < \lambda_0 < \mathrm{vol}(\Omega))$ and $m(\lambda_j) \to \tilde{m}$, then $\tilde{m} = m(\lambda_0)$.

Let u_0, u_j be solutions of (1.12) corresponding to λ_0 and λ_j, respectively. Then

$$m(\lambda_0) = \int_\Omega |\nabla u_0|^2 \qquad \text{and} \qquad m(\lambda_j) = \int_\Omega |\nabla u_j|^2.$$

Since $\Omega \cap \partial\{u_0 > 0\}$ is analytic, the set $\Omega \cap \{u_0 = 0\}$ has finitely many components. Choose subdomain $\tilde{\Omega}_j \subset \Omega$ with

$$\partial\Omega \subset \partial\tilde{\Omega}_j \quad \text{and} \quad \mathrm{vol}(\tilde{\Omega}_j) = \lambda_j = \mathrm{vol}(\{u_j > 0\})$$

in such a way that, if we parametrize $\partial\tilde{\Omega}_j$, $\partial\{u_0 > 0\}$,

$$\begin{aligned} \partial\tilde{\Omega}_j : &\qquad X = X_j(s), &&\qquad s \in \mathcal{I}, \\ \partial\{u_0 > 0\} : &\qquad X = X_0(s), &&\qquad s \in \mathcal{I}, \end{aligned}$$

where $\mathcal{I}$ is a union of finitely many closed intervals, then $X_j \to X_0$ in C^1.

Consider the problem

$$\begin{cases} \triangle \tilde{u}_j = 0 & \text{in} \quad \tilde{\Omega}_j, \\ \tilde{u}_j = 1 & \text{on} \quad \partial\Omega, \\ \tilde{u}_j = 0 & \text{on} \quad \Omega \cap \partial\tilde{\Omega}_j. \end{cases}$$

Extend $\tilde{u}_j$ by 0 into $\Omega \backslash \tilde{\Omega}_j$. Then $\tilde{u}_j \in K_{\lambda_j}$.

It is easy to verify that, for a subsequence,

$$\int_\Omega |\nabla \tilde{u}_j|^2 \to \int_\Omega |\nabla u_0|^2 \qquad \text{as} \quad j \to \infty.$$

Since

$$m(\lambda_j) = \int_\Omega |\nabla u_j|^2 \le \int_\Omega |\nabla \tilde{u}_j|^2,$$

letting $j \to \infty$ gives $\tilde{m} \le m(\lambda_0)$. This result also implies that $m(\lambda_j)$ is bounded.

On the other hand, for a subsequence,

$$\begin{aligned} \nabla u_j &\to \nabla u^* \qquad \text{weakly in } L^2(\Omega), \\ u_j &\to u^* \qquad \text{a.e. in } \Omega. \end{aligned} \tag{2.9}$$

It follows that

$$I_{\{u^*>0\}} \le \liminf_{j\to\infty} I_{\{u_j>0\}} \qquad \text{a.e.},$$

and by Fatou's lemma,

$$\mathrm{vol}(\{u^* > 0\}) \le \liminf_{j\to\infty} \mathrm{vol}(\{u_j > 0\}) = \lim_{j\to\infty} \lambda_j = \lambda_0.$$

Hence $\mathrm{vol}(\Omega) - \lambda_0 \le \mathrm{vol}(\Omega) - \mathrm{vol}(\{u^* > 0\}) = \mathrm{vol}(\{u^* = 0\})$. Therefore we can choose a closed set $D \subset \{u^* = 0\}$ such that $\mathrm{vol}(D) = \mathrm{vol}(\Omega) - \lambda_0$.

Let u be the capacitory potential of the region $\Omega \backslash D$. Then

$$\int_\Omega |\nabla u|^2 \le \int_\Omega |\nabla u^*|^2.$$

Since u is harmonic in $\Omega\backslash D$, $u > 0$ in $\Omega\backslash D$ by the maximum principle. Hence $u \in K_{\lambda_0}$, so that, by (2.9),

$$m(\lambda_0) \leq \int_\Omega |\nabla u|^2 \leq \int_\Omega |\nabla u^*|^2 \leq \liminf \int_\Omega |\nabla u_j|^2 \leq \lim m(\lambda_j) = \tilde{m}.$$

Also, since $\tilde{m} \leq m(\lambda_0)$, $\tilde{m} = m(\lambda_0)$ and the proof of Lemma 2.3 is complete.

LEMMA 2.4. *$m(\lambda)$ is monotonically decreasing.*

Proof. For $0 < \lambda_1 < \lambda_2 < \mathrm{vol}(\Omega)$, let u_1, u_2 be solutions of (1.12) corresponding to λ_1 and λ_2, respectively. Denote by D_i $(i = 1, 2)$ the region $\Omega\backslash\{u_i > 0\}$. Then

$$\mathrm{vol}(D_1) = \mathrm{vol}(\Omega) - \lambda_1 > \mathrm{vol}(\Omega) - \lambda_2 = \mathrm{vol}(D_2).$$

Choose a closed set $\tilde{D} \subset D_1$ such that $\mathrm{vol}(\tilde{D}) = \mathrm{vol}(D_2)$; then $\mathrm{vol}(\Omega\backslash\tilde{D}) = \lambda_2$. By (2.5),

$$m(\lambda_2) = \min_{\mathrm{vol}(\Omega\backslash D)=\lambda_2} \mathrm{Cap}_\Omega D \leq \mathrm{Cap}_\Omega \tilde{D}. \tag{2.10}$$

Since $\mathrm{Cap}_\Omega D$ is monotonically increasing with respect to D,

$$\mathrm{Cap}_\Omega \tilde{D} \leq \mathrm{Cap}_\Omega D_1 = \int_\Omega |\nabla u_1|^2 = m(\lambda_1). \tag{2.11}$$

Combining (2.10) and (2.11), the lemma follows.

LEMMA 2.5. *$m(\lambda) \to 0$ as $\lambda \to \mathrm{vol}(\Omega)$.*

Proof. Let $B_\rho(X_0)$ be the largest ball contained in Ω. By (2.3) and (2.4), $\mathrm{Cap}_\Omega D \leq \mathrm{Cap}_{B_\rho} D$ for any $D \subset B_\rho$.

For λ close to $\mathrm{vol}(\Omega)$, set $\epsilon = \left(\frac{\mathrm{vol}(\Omega)-\lambda}{\pi}\right)^{1/2}$. Then $\mathrm{vol}(\Omega\backslash B_\epsilon(X_0)) = \lambda$, therefore,

$$m(\lambda) \leq \mathrm{Cap}_\Omega B_\epsilon \leq \mathrm{Cap}_{B_\rho} B_\epsilon = \frac{2\pi}{\left|\log \frac{\epsilon}{\rho}\right|}.$$

As $\lambda \to \mathrm{vol}(\Omega)$, we have $\epsilon \to 0$. It follows that $m(\lambda) \to 0$.

LEMMA 2.6. *$m(\lambda) \to \infty$ as $\lambda \to 0+$.*

Proof. Let D_λ be the closed set where solution u_λ vanishes. Then $\mathrm{vol}(D_\lambda) = \mathrm{vol}(\Omega) - \lambda \to \mathrm{vol}(\Omega)$ as $\lambda \to 0$. It follows that there exists a constant $c > 0$ independent of λ such that $\mathrm{vol}(D_\lambda) \geq c$ for all small λ. Since ∂D_λ is analytic, D_λ consists of a finite set of components $\{D_\lambda^i\}$. By the isoperimetric inequality (see [21]),

$$l_{\partial D_\lambda^i} \geq \sqrt{4\pi \mathrm{vol}(D_\lambda^i)},$$

so that

$$l_{\Gamma_\lambda} = l_{\partial D_\lambda} \geq \sum_i \sqrt{4\pi \mathrm{vol}(D_\lambda^i)} \geq \sqrt{4\pi \mathrm{vol}(D_\lambda)} \geq \sqrt{4\pi c} > 0 \qquad \text{for all small } \lambda. \tag{2.12}$$

We shall prove that there exists a sequence $\lambda_j \to 0$ such that

$$c_{\lambda_j} = |\nabla u_{\lambda_j}|(\Gamma_{\lambda_j}) \to \infty. \tag{2.13}$$

Once this has been proved, then by Lemma 2.2 and (2.12), $m(\lambda_j) \to \infty$ as $\lambda_j \to 0$. Hence, by Lemma 2.4,

$$m(\lambda) \to \infty \quad \text{as} \quad \lambda \to 0$$

and the lemma follows.

It remains to prove (2.13). Let $X_0, X_1 \in \partial\Omega$ be such that

$$d = |X_0 - X_1| = \operatorname{diam}(\partial\Omega) \equiv \max_{X,Y\in\partial\Omega} |X - Y|.$$

Since $\operatorname{vol}(D_\lambda) \to \operatorname{vol}(\Omega)$ as $\lambda \to 0$, there exist a sequence $\lambda_j \to 0$ and a set $\{X_j\} \subset \partial D_{\lambda_j}$ such that $X_j \to X_1$.

Consider a sequence of disks $B_{d_j}(X_0)$ containing D_{λ_j} and satisfying

$$\partial B_{d_j} \cap \partial D_{\lambda_j} \neq \emptyset \quad \text{and} \quad \lim_{j\to\infty} d_j = d.$$

Let

$$\phi_j = 1 - \frac{\log \frac{|X-X_0|}{d}}{\log \frac{d_j}{d}}.$$

It is easy to verify that ϕ_j is the solution of the problem

$$\begin{cases} \triangle\phi_j = 0 & \text{in} \quad B_d(X_0)\backslash B_{d_j}(X_0), \\ \phi_j = 1 & \text{on} \quad \partial B_d(X_0), \\ \phi_j = 0 & \text{on} \quad \partial B_{d_j}(X_0). \end{cases}$$

By the maximum principle, $u_{\lambda_j} \geq \phi_j$ and, on $\partial B_{d_j} \cap \partial D_{\lambda_j}$,

$$c_{\lambda_j} = |\nabla u_{\lambda_j}| \geq |\nabla \phi_j| = \frac{1}{d_j |\log \frac{d_j}{d}|} \to \infty \quad \text{as} \quad j \to \infty.$$

This completes the proof of (2.13) and, therefore, Lemma 2.6.

Proof of Theorem 2.1. By Lemmas 2.3–2.6, for any $I > 0$ there exists a unique $0 < \lambda_0 < \operatorname{vol}(\Omega)$ such that $m(\lambda_0) = I$. Denote by u_0 a solution of (1.12) corresponding to λ_0, and denote by Γ_0 the free boundary $\Omega \cap \partial\{u_0 > 0\}$. By Lemma 2.2,

$$\big|\nabla u_0\big|(\Gamma_0) l_{\Gamma_0} = m(\lambda_0) = I,$$

i.e.,

$$\big|\nabla u_0\big| = \frac{I}{l_{\Gamma_0}} \qquad \text{on} \quad \Gamma_0$$

and the proof is complete with $\gamma = \Gamma_0$.

Remark 2.1. Suppose that Γ is convex. By using the geometric variational problem in [3] or [12] instead of (1.12), our existence proof yields a solution for (2.1) with γ convex. Now we shall give a different argument (see [15], p. 339) to prove more, namely, the following theorem.

THEOREM 2.7. *If Γ is convex, then for any solution (γ, u), γ is convex.*

Proof. The function $q = |\nabla u|$ satisfies

$$\begin{aligned} &\triangle q \geq 0 \qquad \text{in} \quad \{u > 0\}, \\ &\frac{\partial q}{\partial \nu} + \kappa q = 0 \qquad \text{along stream lines } \{u = \text{const}\}, \end{aligned}$$

where ν is the normal to the stream line and κ is the curvature of the stream line subject to the sign convention: once ν is chosen, κ is taken as positive when the stream line is convex to the region on the side of the ν direction.

Since q is subharmonic, q does not have interior maximum. Suppose q attains its maximum at $X_0 \in \Gamma$. Take ν as the inward normal. Then $\kappa(X_0) \leq 0$ and

$$0 > \frac{\partial q}{\partial \nu} = -\kappa q \geq 0 \qquad \text{at} \quad X_0,$$

a contradiction. Therefore, $\max q = \max_\gamma q$. Since $q\big|_\gamma = I/l_\gamma = \text{const}$,

$$\kappa\big|_\gamma = -\frac{1}{q}\frac{\partial q}{\partial \nu} > 0 \qquad (\nu = \text{inward normal}),$$

and the assertion follows.

Remark 2.2. Theorems 2.1 and 2.7 can be extended to problem (1.1′), (1.2), (1.3), and (1.4′). Accordingly, functional (1.11) should be replaced by

$$J(v) = \int_\Omega \frac{|\nabla v|^2}{x}.$$

3. Symmetry properties. To begin, we recall some definitions.

We call L a moving line if L is a straight line in $\mathbb{R}^2$ which moves parallel to itself.

Take a moving line L which, initially, does not intersect a set A in $\mathbb{R}^2$. As L moves toward A, it will intersect A and cut off from A a cap $C(A)$ lying behind L.

We define the folding of A (about L) as the reflection of $C(A)$ with respect to L and denote it by A^L. Suppose u is a function defined in A. We define the folding of u (about L) on A^L by

$$u^L(X) = u(X') \qquad \text{for} \quad X' \in C(A), \tag{3.1}$$

where X' is the reflection of X with respect to L.

DEFINITION 3.1. *Suppose L is a straight line in $\mathbb{R}^2$. A curve Γ is called an L-graph if any normal to L which intersects Γ intersects it in either one point or one line segment.*

THEOREM 3.1. *Suppose (γ, u) is a solution of* (2.1). *If Γ is symmetric with respect to a straight line L_0 and the portion of Γ on either side of L_0 is an L_0-graph, then γ is also symmetric about L_0 and the portion of γ on either side of L_0 is an L_0-graph too. Furthermore, $u^{L_0} = u$.*

Proof. We shall use the "folding argument" as in [16] and [22]. Suppose that γ is not symmetric with respect to L_0. We want to derive a contradiction.

Denote by D_γ the bounded domain enclosed by γ. Let L be a moving line in $\mathbb{R}^2$ which is parallel to L_0 and tangent to γ initially. As L moves toward L_0, the folding Γ^L is contained within Γ until $L = L_0$. Obviously, at the early stage of the process, γ^L will be contained in D_γ. Denote by L_e the final position of L, where one of the following events occurs (see Fig. 3.1):

(1) $L_e \neq L_0$, γ^{L_e} becomes internally tangent to γ at some point P not on L_e;

(2) $L_e \neq L_0$, L_e is orthogonal to γ at some Q and γ^{L_e} is contained in D_γ;

(3) $L_e = L_0$, γ^{L_0} does not touch γ.

Denote by $\Omega_\gamma^{L_e}$ the folding of Ω_γ with respect to L_e, which is the domain bounded by Γ^{L_e}, γ^{L_e}, and L_e. Then the folding u^{L_e} of u satisfies the equation

$$\triangle u^{L_e} = 0 \qquad \text{in} \quad \Omega_\gamma^{L_e} \tag{3.2}$$

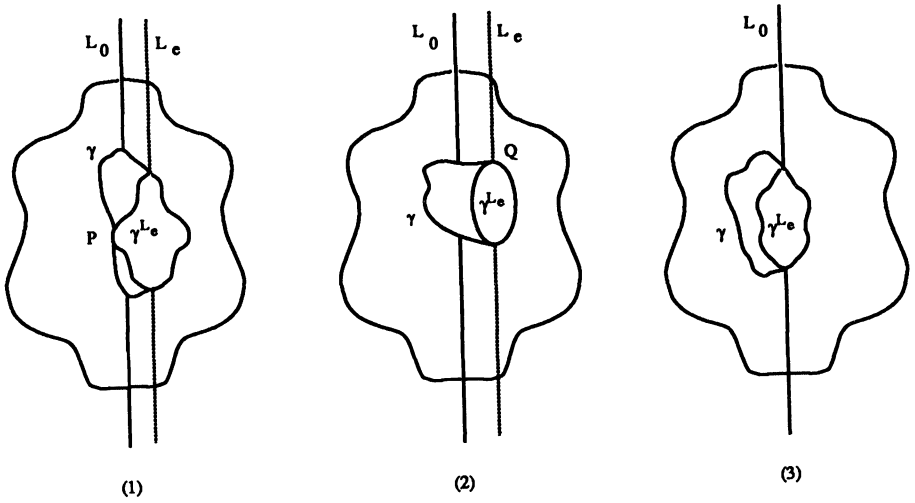


FIG. 3.1.

and the boundary conditions

$$u^{L_e} = 1 \quad \text{on} \quad \Gamma^{L_e}, \tag{3.3}$$

$$u^{L_e} = u \quad \text{on} \quad L_e \cap \Omega_\gamma, \tag{3.4}$$

$$u^{L_e} = 0 \quad \text{on} \quad \gamma^{L_e}, \tag{3.5}$$

$$|\nabla u^{L_e}| = \frac{I}{l_\gamma} \quad \text{on} \quad \gamma^{L_e}. \tag{3.6}$$

Consider the function $v = u^{L_e} - u$ in $\Omega_\gamma \cap \Omega_\gamma^{L_e}$. By our construction, $v \geq 0$ on $\partial(\Omega_\gamma \cap \Omega_\gamma^{L_e})$ and $\triangle v = 0$ in $\Omega_\gamma \cap \Omega_\gamma^{L_e}$. Since $v \not\equiv 0$ in $\Omega_\gamma \cap \Omega_\gamma^{L_e}$, by the strong maximum principle,

$$v > 0 \quad \text{in} \quad \Omega_\gamma \cap \Omega_\gamma^{L_e}. \tag{3.7}$$

We shall prove that (3.7) is impossible.

In case (1), v has its minimum 0 at P. By the maximum principle,

$$\frac{\partial v}{\partial n} > 0 \quad \text{at} \quad P, \qquad n = \text{inward normal},$$

i.e.,

$$\frac{I}{l_\gamma} = |\nabla u^{L_e}| > |\nabla u| = \frac{I}{l_\gamma} \quad \text{at} \quad P.$$

This contradiction shows that case (1) cannot happen.

Suppose we are in case (2). Following the proof for Theorem 1 in [22], we have the following lemma.

LEMMA 3.2. *All the first and second derivatives of v vanish at Q.*

Note that v attains its minimum 0 at Q. By applying the boundary point lemma (Lemma 1) in [22] at Q (which is a right-angled corner of $\Omega_\gamma \cap \Omega_\gamma^{L_e}$) we have, for any inward nontangential direction s,

$$\text{either} \quad \frac{\partial v}{\partial s} > 0 \quad \text{or} \quad \frac{\partial^2 v}{\partial s^2} > 0 \quad \text{at} \quad Q,$$

contradicting Lemma 3.2. Therefore case (2) is also impossible.

Finally, if we are in case (3), then L_0 divides D_γ into two parts D_γ^1 and D_γ^2 with $\mathrm{vol}(D_\gamma^1) > \mathrm{vol}(D_\gamma^2)$. Let $L \neq L_0$ be a straight line parallel to L_0 such that L lies on the same side of L_0 as D_γ^1 and $L \cap D_\gamma^1 = \emptyset$. As L moves parallel to itself toward L_0, one of the events (1) or (2) will occur. However, this is impossible and γ must be symmetric with respect to L_0.

The portion of γ on either side of L_0 must be an L_0-graph. Otherwise, before L reaches L_0, one of the cases (1) or (2) will occur, but these cases have already been ruled out.

Remark 3.1. By uniqueness to the Cauchy problem and unique continuation for solutions of elliptic equations, γ cannot contain a line segment.

Remark 3.2. Theorem 3.1 can be extended to high dimensions without any change of the proof.

COROLLARY 3.3. *If Γ is an n-sphere with radius R, then* (2.1) *has a unique solution (γ, u) with γ as a concentric n-sphere having radius*

$$R_0 = \begin{cases} Re^{-\frac{2\pi}{I}}, & n = 2, \\ \dfrac{1}{\left[\frac{(n-2)\omega_{n-1}}{I} + \frac{1}{R^{n-2}}\right]^{\frac{1}{n-2}}}, & n > 2, \end{cases} \tag{3.8}$$

where ω_{n-1} is the surface area of a unit n-sphere and

$$u = \begin{cases} \left(1 + \frac{I}{2\pi}\log\frac{r}{R}\right), & R_0 < r < R, \quad n = 2, \\ \left[1 - \frac{I}{(n-2)\omega_{n-1}}\left(\frac{1}{r^{n-2}} - \frac{1}{R^{n-2}}\right)\right], & R_0 < r < R, \quad n > 2, \end{cases} \tag{3.9}$$

where r denotes distance from the center.

Proof. Suppose (γ, u) is a solution. By Theorem 3.1 and Remark 3.2, γ is a concentric sphere, which in turn implies that u is radial. A simple computation shows that (γ, u) must have the form (3.8), (3.9).

4. Uniqueness. Suppose Γ is a $C^{1+\alpha}$ $(0 < \alpha < 1)$ plane curve. Denote by Γ^+ the portion of Γ lying in the first quadrant. We assume that

(4.1)
(1) Γ is symmetric with respect to the x-axis and the y-axis;
(2) Γ^+ is both an x-graph and a y-graph.

Denote by (γ, u) a solution of (2.1). From Theorem 3.1 it follows that γ is symmetric with respect to the two coordinate axes and γ^+ is both an x-graph and a y-graph.

LEMMA 4.1. *$xu_x \geq 0$ and $yu_y \geq 0$ in Ω_γ.*

Proof. Since $\Gamma \cap \{x > 0\}$ and $\gamma \cap \{x > 0\}$ are both y-graphs and u takes its maximum 1 on $\Gamma \cap \{x > 0\}$ and minimum 0 on $\gamma \cap \{x > 0\}$, we have

$$u_x \geq 0 \quad \text{on} \quad \Gamma \cap \{x > 0\} \cup \gamma \cap \{x > 0\}. \tag{4.2}$$

By Theorem 3.1, u is an even function with respect to x, i.e.,

$$u_x\big|_{x=0} = 0. \tag{4.3}$$

Combining (4.2) and (4.3), we get

$$u_x \geq 0 \quad \text{on} \quad \partial\{\Omega_\gamma \cap \{x > 0\}\}.$$

Since u_x is harmonic in $\Omega_\gamma \cap \{x > 0\}$, by the maximum principle we have

$$u_x \geq 0 \qquad \text{in} \quad \Omega_\gamma \cap \{x > 0\}.$$

Since u is even in x, $u_x \leq 0$ in $\Omega_\gamma \cap \{x < 0\}$ and we conclude that

$$xu_x \geq 0 \qquad \text{in} \quad \Omega_\gamma.$$

The same argument applied to u_y shows that

$$yu_y \geq 0 \qquad \text{in} \quad \Omega_\gamma,$$

and the proof is complete.

THEOREM 4.2. *Under assumption* (4.1), *for any* $I > 0$ *there exists at most one solution of the problem* (2.1).

Proof. Suppose there are two solutions (γ_i, u_i) $(i = 1, 2)$ with $\gamma_1 \not\equiv \gamma_2$. We shall derive a contradiction.

Denote by D_{γ_i} the bounded domain enclosed by γ_i. By symmetry, $D_{\gamma_1} \cap D_{\gamma_2} \neq \emptyset$. We shall show that neither $D_{\gamma_1} \subset D_{\gamma_2}$ nor $D_{\gamma_2} \subset D_{\gamma_1}$.

Suppose $D_{\gamma_1} \subset D_{\gamma_2}$. Then, by the strong maximum principle,

$$u_2 < u_1 \qquad \text{in} \quad \Omega_{\gamma_2}.$$

Since $u_1 = u_2 = 1$ on Γ, by the maximum principle we have

$$\frac{\partial u_2}{\partial n} > \frac{\partial u_1}{\partial n} \qquad \text{on} \quad \Gamma \quad (n = \text{outward normal}). \tag{4.4}$$

On the other hand, using the boundary condition $-\partial u_i/\partial n = |\nabla u_i| = I/l_{\gamma_i}$ on γ_i,

$$\int_\Gamma \frac{\partial u_i}{\partial n} = -\int_{\gamma_i} \frac{\partial u_i}{\partial n} = I.$$

Hence

$$\int_\Gamma \frac{\partial u_1}{\partial n} ds = \int_\Gamma \frac{\partial u_2}{\partial n} ds = I, \tag{4.5}$$

contradicting (4.4).

The same argument shows that $D_{\gamma_2} \subset D_{\gamma_1}$ is also impossible. Consequently, γ_1 and γ_2 must intersect. Since γ_1 and γ_2 are analytic, they intersect only at a finite number of points. Take such a point Q. Then, for each neighborhood N of Q, $N \cap \gamma_i$ intersects both D_{γ_j} and $D^c_{\gamma_j}$, i.e.,

$$N \cap \gamma_i \cap D_{\gamma_j} \neq \emptyset \quad \text{and} \quad N \cap \gamma_i \cap D^c_{\gamma_j} \neq \emptyset \quad (i \neq j; i, j = 1, 2).$$

Denote by $T(\gamma_1 \cap \gamma_2)$ the set of all points of intersection of γ_1 and γ_2. By symmetry, $T(\gamma_1^+ \cap \gamma_2^+) \neq \emptyset$, where $\gamma_i^+ (i = 1, 2)$ denotes the portion of γ_i lying in the first quadrant.

We first consider the case where

$$T(\gamma_1^+ \cap \gamma_2^+) \text{ contains a single point } Q. \tag{4.6}$$

Set $w = u_1 - u_2$ in $\Omega_{\gamma_1} \cap \Omega_{\gamma_2}$. By symmetry and (4.5), there exists a point $X^+ \in \Gamma^+$ where $\frac{\partial w}{\partial n} = 0$, i.e., $|\nabla w| = 0$. It follows that there exists a level curve $\sigma_1 \subset \{w = 0\}$ starting at X^+ and going into $\Omega_{\gamma_1} \cap \Omega_{\gamma_2}$. To see this, let g be the conformal mapping which maps the domain $B_\rho(X^+) \cap \Omega_{\gamma_1} \cap \Omega_{\gamma_2}$ into $\{\tilde{y} > 0\}$, $B_\rho(X^+) \cap \Gamma^+$ into $\{\tilde{y} = 0\}$,

and $g(X^+) = 0$, where $B_\rho(X^+)$ is a disc with the center at X^+ and the radius ρ being a small positive number. Since $w = 0$ on Γ^+, we define a harmonic function v in a small disc $B_r(0)$ by

$$v(\tilde{x}, \tilde{y}) = \begin{cases} w(g^{-1}(\tilde{x}, \tilde{y})), & (\tilde{x}, \tilde{y}) \in B_r(0) \cap \{\tilde{y} > 0\}, \\ -w(g^{-1}(\tilde{x}, -\tilde{y})), & (\tilde{x}, \tilde{y}) \in B_r(0) \cap \{\tilde{y} < 0\}. \end{cases}$$

Since $w(X^+) = 0$, $|\nabla w(X^+)| = 0$, we have

$$v(0) = 0, \qquad |\nabla v(0)| = 0.$$

It follows that 0 is the branch point of the level curves $\{v = 0\}$. Choose a branch $\tau \subset \{v = 0\}$ initiating from 0 and going into $\{\tilde{y} > 0\}$. Then $\sigma_1 = g^{-1}(\tau)$ is a level curve of $\{w = 0\}$ starting at X^+ and going into $\Omega_{\gamma_1} \cap \Omega_{\gamma_2} \cap \{x > 0, y > 0\}$.

Denote by σ_2, σ_3, and σ_4, respectively, the symmetric counterparts of σ_1 in the second, third, and fourth quadrants.

σ_1 cannot exit the first quadrant on the y-axis. Suppose otherwise and denote by $\Omega_0 \subset \Omega$ the domain bounded by σ_1, σ_2, and Γ. Then $w = 0$ on $\partial\Omega_0$, which implies, by the maximum principle, that $w \equiv 0$ in Ω_0. By unique continuation, $w = 0$ everywhere and this is impossible by our assumption that $\gamma_1 \not\equiv \gamma_2$. Similarly, σ_1 cannot end either on the x-axis or Γ^+. Hence σ_1 must terminate at a point P in $\gamma_1^+ \cap \gamma_2^+$ with $P = (x_p, y_p)$, $x_p > 0$, $y_p > 0$ (P is not necessarily a point in $T(\gamma_1^+ \cap \gamma_2^+)$).

Consider the cases $P = Q$ and $P \neq Q$ separately.

(i) $P = Q$. Denote by $\gamma_i(P) = \gamma_i \cap \{-x_p < x < x_p, y > 0\}$ $(i = 1, 2)$ the portion of γ_i in the upper half plane lying between $\sigma_1 \cap \gamma_i$ and $\sigma_2 \cap \gamma_i$. Since one $\gamma_i(P)$ lies outside the domain bounded by the other free boundary curve, we may assume that (see Fig. 4.1)

$$\gamma_1(P) \quad \text{lies outside} \quad D_{\gamma_2}. \tag{4.7}$$

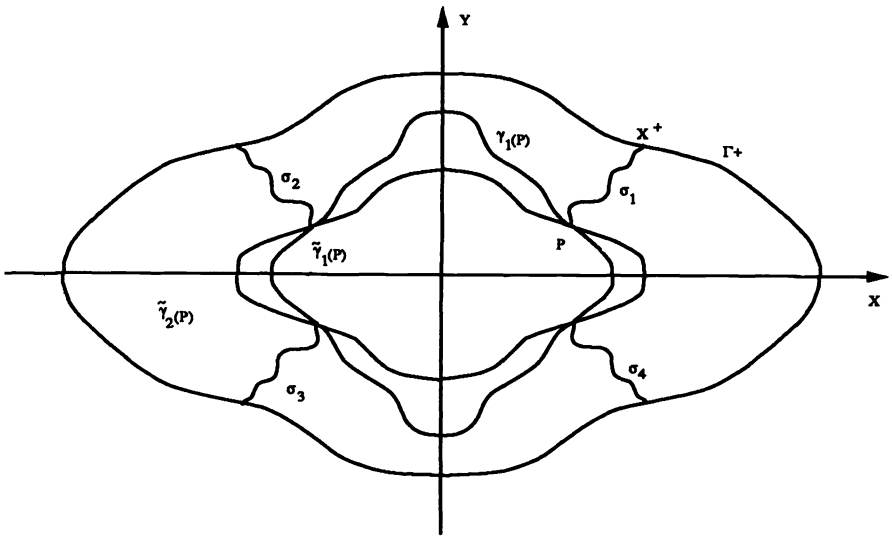


FIG. 4.1.

Denote by $\tilde{\gamma}_i(P) = \gamma_i \cap \{x < 0, -y_p < y < y_p\}$ the portion of γ_i in the left half

plane lying between $\sigma_2 \cap \gamma_i$ and $\sigma_3 \cap \gamma_i$. Then, by (4.6) and (4.7) (see Fig. 4.1),

$$\tilde{\gamma}_2(P) \quad \text{lies outside} \quad D_{\gamma_1}. \tag{4.8}$$

Take a vertical line L in $\{x > 0\}$ which is tangent to Γ initially and move it in a parallel fashion toward the y-axis. We fold Ω_{γ_2} with respect to L. Denote by $\Omega_{\gamma_2}^L$, Γ^L, and γ_2^L the folding of Ω_{γ_2}, Γ, and γ_2, respectively. By (4.7) and (4.8), L will reach the position at $\{x = l > 0\}$, so that $\gamma_2^L \cap \{x < 0\}$ lies inside D_{γ_1} and is tangent to γ_1 at $X_0 = (x_0, y_0)$. We shall prove that $X_0 \in \tilde{\gamma}_1(P)$, i.e., it does not lie on $\gamma_1(P)$.

Clearly, $x_0 < 0$. By symmetry, we may assume that $y_0 \geq 0$. Since γ_2^+ is both an x-graph and a y-graph, γ_2^L is contained in D_{γ_2}; it follows by (4.7) that γ_2^L does not touch $\gamma_1(P)$, which implies that $y_0 < y_p$. Therefore

$$X_0 \in \gamma_2^L \cap \tilde{\gamma}_1(P) \tag{4.9}$$

and

$$\operatorname{dist}(X_0, \Gamma^L) \geq \operatorname{dist}(\gamma_2^L, \Gamma^L) \geq \operatorname{dist}(\gamma_2, \Gamma) > 0. \tag{4.10}$$

In order to apply the maximum principle to the function $u_2^L - u_1$ (u_2^L is the folding of u_2), we shall construct a domain $\Omega^* \subset \Omega_{\gamma_2}^L \cap \Omega_{\gamma_1}$ such that X_0 belongs to a smooth portion of $\partial\Omega^*$. Since $\Gamma^L \cap \tilde{\gamma}_1(P)$ may be nonempty, $\tilde{\gamma}_1(P)$ would be cut into several pieces by the points in $\Gamma^L \cap \tilde{\gamma}_1(P)$. However, by (4.9) and (4.10), there exists a connected portion $\gamma_1^* \subset \tilde{\gamma}_1(P) \backslash \{\Gamma^L \cap \tilde{\gamma}_1(P)\}$ which contains X_0 in its interior and has two end points P_1, P_2 such that one of the following cases happens:

(a) $P_1, P_2 \in \Gamma^L \cap \tilde{\gamma}_1(P)$;

(b) $P_1 \in \Gamma^L \cap \tilde{\gamma}_1(P)$, $P_2 \in \gamma_1 \cap \sigma_2$;

(c) $P_1 \in \gamma_1 \cap \sigma_3$, $P_2 \in \gamma_1 \cap \sigma_2$.

Since Γ^L is contained in the domain enclosed by Γ, $\Gamma^L \cap \sigma_i \neq \emptyset$ $(i = 2, 3)$. Define (see Fig. 4.2)

$$\Omega^* = \begin{cases} \text{the domain bounded by } \Gamma^L,\ \gamma_1^* & \text{if (a) occurs,} \\ \text{the domain bounded by } \Gamma^L,\ \gamma_1^*, \text{ and } \sigma_2 & \text{if (b) occurs,} \\ \text{the domain bounded by } \Gamma^L,\ \gamma_1^*,\ \sigma_2,\ \sigma_3 & \text{if (c) occurs.} \end{cases} \tag{4.11}$$

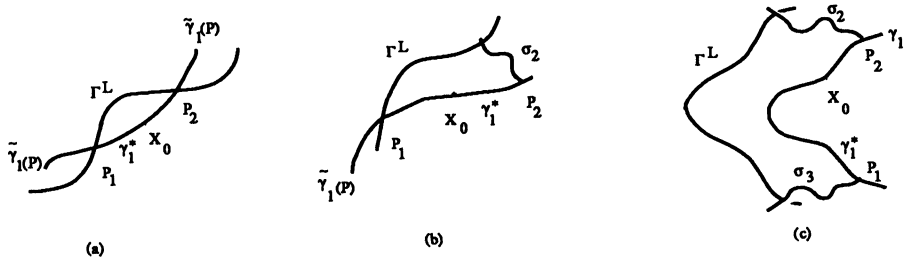


FIG. 4.2.

We now look at the boundary values of u_2^L and u_1 on $\partial\Omega^*$. Since $u_2^L = 1$ on Γ^L and $u_1 = 0$ on γ_1^*, we have

$$u_2^L \geq u_1 \quad \text{on} \quad \Gamma^L \quad \text{and} \quad \gamma_1^*. \tag{4.12}$$

This inequality also holds on $\sigma_2 \cup \sigma_3$.

For any $X \in \sigma_2 \cup \sigma_3$,

$$\begin{aligned} u_2^L(X) = u_2(2l - x, y) &\geq u_2(-x, y) \qquad \text{(by Lemma 4.1)} \\ &= u_2(x, y) \qquad \text{(by symmetry)} \\ &= u_1(X) \qquad \text{(by the definition of } \sigma_2 \text{ and } \sigma_3). \end{aligned}$$

Hence

$$u_2^L \geq u_1 \qquad \text{on} \quad \partial\Omega^*. \tag{4.13}$$

By the strong maximum principle it follows that

$$u_2^L > u_1 \qquad \text{in} \quad \Omega^*.$$

Since $u_2^L = u_1 = 0$ at X_0, applying the maximum principle we get

$$|\nabla u_2^L| > |\nabla u_1| \qquad \text{at} \quad X_0,$$

i.e.,

$$\frac{I}{l_{\gamma_2}} = |\nabla u_2| = |\nabla u_2^L| > |\nabla u_1| = \frac{I}{l_{\gamma_1}} \qquad \text{at} \quad X_0,$$

which implies that

$$l_{\gamma_1} > l_{\gamma_2}. \tag{4.14}$$

Note that (4.7) and (4.8) can be interchanged by index permutation $1 \leftrightarrow 2$ and the coordinate transformation $(x, y) \to (-y, x)$, which preserves harmonicity and boundary conditions of u_i. Hence the preceding argument implies that

$$l_{\gamma_2} > l_{\gamma_1}. \tag{4.15}$$

This, however, contradicts (4.14), and the proof for case (i) is complete.

(ii) $P \neq Q$. The point P cuts γ_i^+ $(i = 1, 2)$ into two pieces, one of which contains Q. Denote by $\hat{\gamma}_i(P) = \gamma_i \cap \{x > 0, -y_p < y < y_p\}$ the portion of γ_i in the right half plane which lies between $\sigma_1 \cap \gamma_i$ and $\sigma_4 \cap \gamma_i$. Then

$$\text{either} \quad Q \in T(\hat{\gamma}_1(P) \cap \hat{\gamma}_2(P)) \quad \text{or} \quad Q \in T(\gamma_1(P) \cap \gamma_2(P)).$$

Rotating the coordinate if necessary, we may assume that (see Fig. 4.3)

$$Q \in T(\hat{\gamma}_1(P) \cap \hat{\gamma}_2(P)). \tag{4.16}$$

Condition (4.16) means that $\hat{\gamma}_i(P)$ intersects both D_{γ_j} and $D_{\gamma_j}^c$ $(i \neq j; i, j = 1, 2)$.

Using the folding argument as before, we assert that there is a vertical line L in $\{x > 0\}$ such that $\hat{\gamma}_1(P)^L \cap \{x < 0\}$ lies inside D_{γ_2} and is tangent to γ_2 at a point $X_0 = (x_0, y_0)$ with $x_0 < 0$, $y_0 \geq 0$. We shall argue that

$$X_0 \in \tilde{\gamma}_2(P) = \gamma_2 \cap \{x < 0, -y_p < y < y_p\}. \tag{4.17}$$

Since γ_i^+ (i=1,2) are both x-graphs and y-graphs, $\hat{\gamma}_1(P)^L$ lies below the segment τ connecting P and $P' = (-x_p, y_p)$. It follows that $\hat{\gamma}_1(P)^L$ does not touch the portion of γ_2 lying above τ which is precisely $\gamma_2(P)$. Hence $y_0 < y_p$ and (4.17) follows.

Therefore a domain Ω^* can be constructed in the same way as (4.11) so that X_0 is contained in a smooth portion of $\partial\Omega^*$ and

$$u_1^L \geq u_2 \qquad \text{on} \quad \partial\Omega^*.$$

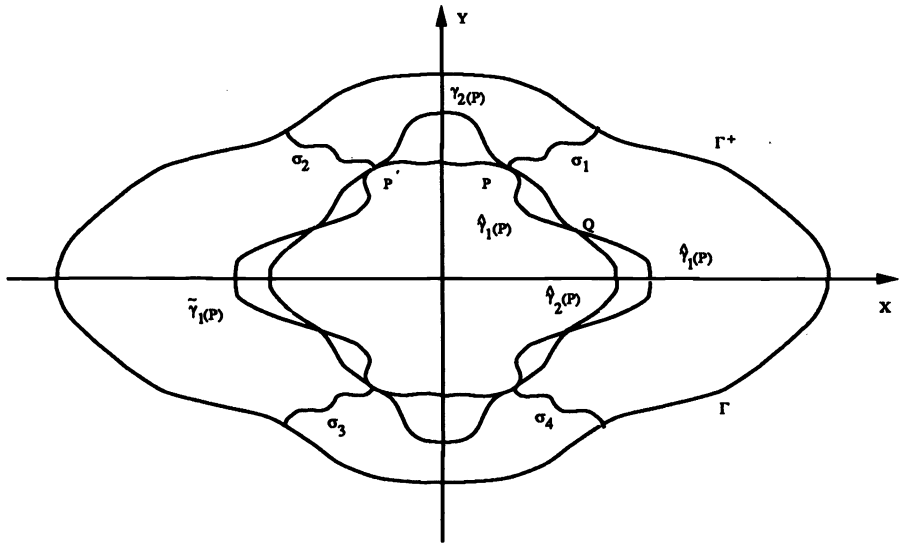


FIG. 4.3

By the strong maximum principle, $u_1^L > u_2$ in Ω^* and

$$|\nabla u_1^L| > |\nabla u_2| \qquad \text{at} \quad X_0,$$

i.e.,

$$\frac{I}{l_{\gamma_1}} = |\nabla u_1| = |\nabla u_1^L| > |\nabla u_2| = \frac{I}{l_{\gamma_2}} \qquad \text{at} \quad X_0,$$

which implies that

$$l_{\gamma_2} > l_{\gamma_1}. \tag{4.18}$$

Since we accomplished the above argument under assumption (4.16), which is symmetric with respect to the indices 1 and 2, we also have

$$l_{\gamma_1} > l_{\gamma_2}, \tag{4.19}$$

a contradiction. This completes the proof of uniqueness for case (4.6).

It remains to consider the case where

$$\text{the set } T(\gamma_1^+ \cap \gamma_2^+) \text{ contains at least 2 points.} \tag{4.20}$$

Denote by $\sigma_1 \subset \{u_1 = u_2\}$ a level curve which initiates at a point $X^+ \in \Gamma^+$ and terminates at a point $P \in \gamma_1^+ \cap \gamma_2^+$. By (4.20) there exists a point $Q \in T(\gamma_1^+ \cap \gamma_2^+)$ such that $P \neq Q$.

We adopt the same notation $\hat{\gamma}_i(P)$, $\gamma_i(P)$, etc. Without loss of generality, we may assume that (see Fig. 4.4)

$$Q \in T(\hat{\gamma}_1(P) \cap \hat{\gamma}_2(P))$$

(this assumption is the same as (4.16)). Recalling the arguments which follow (4.16) and (4.17), the folding $\hat{\gamma}_i(P)^L$ does not touch $\gamma_j(P)$ no matter what happens on $\gamma_j(P)$. Therefore we can reach the same situation as (4.17) by the folding argument and construct a domain Ω^* with X_0 contained in a smooth portion of $\partial\Omega^*$. Finally,

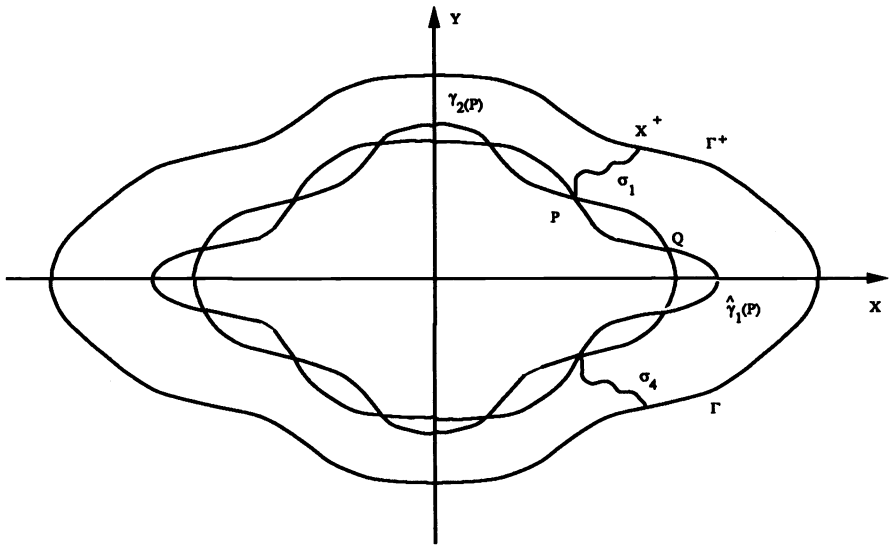


FIG. 4.4

by the maximum principle applied to $u_i^L - u_j$ $(i \neq j)$, we can derive a contradiction in the same way as before. This completes the proof for Theorem 4.2.

COROLLARY 4.3. *Consider the variational problem* (1.12). *If* $\partial\Omega$ *satisfies conditions* (1) *and* (2) *of* (4.1), *then* (1.12) *has a unique solution.*

Proof. By results in §2, any minimizer of (1.12) solves (2.1) for $I = m^{-1}(\lambda)$. It follows from Theorem 4.2 that the solution of (1.12) is unique.

Remark 4.1. Examples in the next section will show that condition (2) is essential for assuring the uniqueness of (2.1): any small perturbation may result in nonuniqueness.

5. Nonuniqueness. Take a unit disc B_1 centered at the origin and two discs B_ϵ of radius $\epsilon << 1$, one centered at $(1 + 2\epsilon, 0)$ and another centered at $(-1 - 2\epsilon, 0)$. Connect the three discs by two "thin corridors" T_δ with length ϵ and width $\delta < \epsilon$ to form a domain Ω with a smooth boundary which is symmetric with respect to both coordinate axes. We regard Ω as a small perturbation of B_1.

THEOREM 5.1. *If* δ *is sufficiently small, then problem* (2.1) *has at least two distinct solutions, one of which is symmetric with respect to the two coordinate axes and another one which is nonsymmetric.*

We first construct a nonsymmetric solution using a variational formulation with an obstacle.

Choose a smooth function ϕ (the "obstacle") depending only on x such that

$$
\begin{aligned}
&\phi'' \geq 0 && \text{in } \Omega, \\
&0 \leq \phi \leq \frac{1}{2} && \text{in } \Omega, \\
&\phi > 0 && \text{in } \Omega \cap \{x < x_0\}, \quad \text{and} \\
&\phi \equiv 0 && \text{in } \Omega \cap \{x > x_0\},
\end{aligned}
\tag{5.1}
$$

where $1 + \frac{1}{4}\epsilon < x_0 < 1 + \frac{1}{2}\epsilon$, for example,

$$\phi = a((x - x_0)^-)^4,$$

where a is a small positive constant.

For $0 < \lambda < \mathrm{vol}(\Omega \cap \{x > x_0\})$, set

$$K_{\phi,\lambda} = \{v \in H^1(\Omega);\quad v(.,y) = v(.,-y), v \geq \phi, \quad v\big|_{\partial\Omega} = 1 \quad \text{and} \quad \mathrm{vol}(\{v = 0\}) = \lambda\}. \tag{5.2}$$

Consider the following problem: find $u \in K_{\phi,\lambda}$ such that

$$J(u) = \min_{v \in K_{\phi,\lambda}} J(v), \qquad J(v) = \int_\Omega |\nabla v|^2. \tag{5.3}$$

The reason for introducing the obstacle ϕ is to prevent free boundary $\Omega \cap \partial\{u > 0\}$ from appearing in $\Omega \cap \{x < x_0\}$.

To study problem (5.3), we introduce a related penalized problem.

For $\eta > 0$, introduce the functional

$$J_\eta(v) = J(v) - f_\eta(\mathrm{vol}(\{v = 0\})), \tag{5.4}$$

where

$$f_\eta(s) = \begin{cases} \eta(s - \lambda) & \text{for} \quad s \geq \lambda, \\ \frac{1}{\eta}(s - \lambda) & \text{for} \quad s \leq \lambda \end{cases} \tag{5.5}$$

and the admissible class

$$K_\phi = \{v \in H^1(\Omega);\quad v(.,y) = v(.,-y), v \geq \phi, v\big|_{\partial\Omega} = 1\}. \tag{5.6}$$

Consider the following problem: find $u \in K_\phi$ such that

$$J_\eta(u) = \min_{v \in K_\phi} J_\eta(v). \tag{5.7}$$

By the standard argument (see [7]), there exists a solution u_η of (5.7). Since its increasing rearrangement u_η^* with respect to $\{y = 0\}$ is also in K_ϕ and (see [21])

$$\mathrm{vol}(\{u_\eta^* = 0\}) = \mathrm{vol}(\{u_\eta = 0\}), \quad \int_\Omega |\nabla u_\eta^*|^2 \leq \int_\Omega |\nabla u_\eta|^2,$$

u_η^* is also a minimizer of (5.7). In what follows we use u to denote such a minimizer; it satisfies $u_y \geq 0$ for $y \geq 0$.

LEMMA 5.2. *For every* disc $B_r(X^0) \subset \Omega$,

$$\int_{B_r(X^0)} |\nabla(u - v)|^2 \leq \frac{2}{\eta} \int_{B_r(X^0)} I_{\{u=0\}},$$

where v is the harmonic function in $B_r(X^0)$ taking boundary value u on $\partial B_r(X^0)$.

Proof. If either $B_r(X^0) \subset \Omega \backslash \{y = 0\}$ or $X^0 = (x^0, 0)$, the proof follows from [15, p. 276]. It only remains to consider the complementary case where

$$B_r(X^0) \text{ intersects } \{y = 0\} \text{ and is nonsymmetric with respect to } \{y = 0\}.$$

Denote by B the union of $B_r(X^0) \cap \{y \geq 0\}$ with its reflection about $\{y = 0\}$. Let $\tilde{v}$ be the function which is harmonic in B and equal to u in $\Omega \backslash \bar{B}$. Since ϕ is subharmonic, $\tilde{v} \geq \phi$ such that $\tilde{v} \in K_\phi$. Developing $J_\eta(u) \leq J_\eta(\tilde{v})$ and using the fact

that $\eta \le f'_\eta \le \frac{1}{\eta}$, we have

$$\int_B |\nabla(u-\tilde{v})|^2 \le \frac{1}{\eta}\int_B I_{\{u=0\}} = \frac{2}{\eta}\int_{B\cap\{y>0\}} I_{\{u=0\}} \le \frac{2}{\eta}\int_{B_r(X^0)} I_{\{u=0\}}. \tag{5.8}$$

Suppose v is harmonic in $B_r(X^0)$, taking the boundary value u on $\partial B_r(X^0)$. Extend v by u into B. Since $\tilde{v}$ is harmonic in B and equal to $v(=u)$ on ∂B,

$$\int_B |\nabla\tilde{v}|^2 \le \int_B |\nabla v|^2,$$

which implies that

$$\int_B |\nabla u|^2 - \int_B |\nabla v|^2 \le \int_B |\nabla u|^2 - \int_B |\nabla\tilde{v}|^2 = \int_B |\nabla(u-\tilde{v})|^2. \tag{5.9}$$

Combining (5.8) and (5.9), we have

$$\int_{B_r(X^0)} |\nabla(u-v)|^2 = \int_B |\nabla u|^2 - \int_B |\nabla v|^2 \le \frac{2}{\eta}\int_{B_r(X^0)} I_{\{u=0\}}.$$

The lemma follows.

Proceeding as in [15, Chap. 3, Lems. 3.1 and 3.2], one can establish that

$$u \in C^{0,1}(\Omega). \tag{5.10}$$

For $C^\alpha(0<\alpha<1)$ regularity we refer to [19].

Since u is continuous by (5.10), one can easily prove that u is harmonic in $\Omega\cap\{u>\phi\}$. However, a stronger result is given by the following lemma.

LEMMA 5.3. *u is harmonic in $\Omega\cap\{u>0\}$.*

Proof. For any disc $B\subset\Omega\cap\{u>0\}$, let v be the harmonic function in B taking boundary value u on ∂B. By Lemma 5.2,

$$\int_B |\nabla(u-v)|^2 \le \frac{2}{\eta}\int_B I_{\{u=0\}} = 0.$$

Hence $u\equiv v$ in B and u is harmonic in B. It follows that u is harmonic in $\Omega\cap\{u>0\}$.

Denote by Γ_η the free boundary $\Omega\cap\partial\{u>0\}$; Γ_η may touch the set $\{x=x_0\}$ (x_0 is given in (5.1)).

LEMMA 5.4. *$\Gamma_\eta\cap\{x>x_0\}$ is analytic and there exists a continuous function $f(x)\ge 0$, $x\in\mathcal{I}$ (an open subset of $\{x_0<x<1+3\epsilon\}$) such that*

$$\Gamma_\eta\cap\{x>x_0\} = \{(x,y);\quad y=f(x), x\in\mathcal{I}\}\cup\{(x,y);\quad y=-f(x), x\in\mathcal{I}\}.$$

Moreover,

$$|\nabla u| = c_\eta \qquad \text{on}\quad \Gamma_\eta\cap\{x>x_0\}, \tag{5.11}$$

where c_η is an unprescribed constant.

Proof. The proof is similar to that in [7]. The only difference comes from the presence of the obstacle in our variational formulation. However, in the region $\Omega\cap\{x>x_0\}$, the obstacle vanishes. Therefore the results in [7] apply to the present situation, i.e., $\Gamma_\eta\cap\{x>x_0\}$ is an analytic curve and u satisfies (5.11). Since $u_y\ge 0$ for $y\ge 0$,

and $u(x,y) = u(x,-y)$, there exists a function $f(x) \geq 0$, $x \in \mathcal{I} \subset \{x_0 < x < 1+3\epsilon\}$ such that

$$\{u = 0\} = \{-f(x) \leq y \leq f(x), \quad x \in \mathcal{I}\}.$$

By uniqueness for the Cauchy problem and unique continuation for solutions to elliptic equations, $\Gamma_\eta \cap \{x > x_0\}$ does not contain a line segment. Therefore f is continuous in $\mathcal{I}$ and the proof of the lemma is complete.

Remark 5.1. $\mathcal{I}$ is the union of intervals in $\{x_0 < x < 1+3\epsilon\}$. By the nonoscillation lemma in [15, p. 287], f is continuous up to the end points of the intervals.

Next we show that if η is small, then $\text{vol}(\{u = 0\}) = \lambda$ so that by (5.4) and (5.5), u is a solution of (5.3).

In what follows let c and C denote any positive constants independent of η. Assume that

$$0 < \lambda < \text{vol}(\Omega \cap \{x > x_0\}). \tag{5.12}$$

Then we can construct a function $\tilde{u} \in K_\phi$ with $\text{vol}(\{\tilde{u} = 0\}) = \lambda$. Since u is a minimizer,

$$J_\eta(u) \leq J_\eta(\tilde{u}) = \int_\Omega |\nabla \tilde{u}|^2 \leq C.$$

It follows that

$$\int_\Omega |\nabla u|^2 \leq C + f_\eta(\text{vol}(\{u = 0\})) \leq C + \eta(\text{vol}(\Omega) - \lambda) \leq C \quad (\eta \text{ small}) \tag{5.13}$$

and

$$-f_\eta(\text{vol}(\{u = 0\})) \leq C,$$

i.e.,

$$\begin{cases} -\eta(\text{vol}(\{u = 0\}) - \lambda) \leq C & \text{if } \text{vol}(\{u = 0\}) \geq \lambda, \\ -\frac{1}{\eta}(\text{vol}(\{u = 0\}) - \lambda) \leq C & \text{if } \text{vol}(\{u = 0\}) \leq \lambda. \end{cases}$$

Combining the two cases, we have

$$\text{vol}(\{u = 0\}) \geq \lambda - C\eta \geq \frac{\lambda}{2} \quad \text{for all } \eta \text{ small.} \tag{5.14}$$

LEMMA 5.5. *Suppose δ is chosen so that $\delta < \sqrt{2\pi\lambda}$ and $\text{vol}(T_\delta) \leq \frac{\lambda}{4}$. Then $c \leq c_\eta \leq C$ for all η small (c_η is the constant in* (5.11)).

Proof. By (5.13), Lemma 5.3, and integration by parts,

$$\begin{aligned} C \geq \int_\Omega |\nabla u|^2 &= \int_{\partial\Omega} \frac{\partial u}{\partial n} ds \\ &= -\int_{\Omega \cap \partial\{u>0\}} \frac{\partial u}{\partial n} ds \qquad (n = \text{outward normal}) \\ &= \int_{\Gamma_\eta \cap \{x = x_0\}} |\nabla u| ds + \int_{\Gamma_\eta \cap \{x > x_0\}} |\nabla u| ds \\ &\geq c_\eta l_{\Gamma_\eta \cap \{x > x_0\}} \qquad \text{(by (5.11)).} \end{aligned}$$

Using the isoperimetric inequality (2.12), we have

$$l_{\Gamma_\eta} \geq \sqrt{4\pi \text{vol}(\Omega \cap \{u = 0\})} \geq \sqrt{2\pi\lambda} \qquad \text{by (5.14).}$$

Since $\Gamma_\eta \cap \{x = x_0\} \leq \delta$,

$$l_{\Gamma_\eta \cap \{x > x_0\}} \geq \sqrt{2\pi\lambda} - \delta.$$

It follows that

$$c_\eta \leq \frac{C}{\sqrt{2\pi\lambda} - \delta} \leq C.$$

It remains to estimate c_η from below. Denote by B_ϵ the disc lying in $\{x > 0\}$ and constituting a portion of Ω. By (5.14) and the assumption $\mathrm{vol}(T_\delta) \leq \frac{\lambda}{4}$,

$$\mathrm{vol}(B_\epsilon \cap \{u = 0\}) = \mathrm{vol}(\{u = 0\}) - \mathrm{vol}(T_\delta \cap \{u = 0\}) \geq \frac{\lambda}{4} \tag{5.15}$$

for all small η. Take a disc B_r with r independent of η such that $\{u = 0\} \subset B_r$. Since $B_\epsilon \cap \Gamma_\eta \neq \emptyset$ (by (5.15)), we can translate B_r until ∂B_r touches $\Gamma_\eta \cap \{x > x_0\}$. Denote by X_η the center of the disc B_r.

Choose a larger disc $B_R(X_\eta)$ containing Ω and consider the auxiliary function

$$\phi_\eta(X) = 1 - \frac{\log \frac{|X - X_\eta|}{R}}{\log \frac{r}{R}}, \qquad X \in B_R(X_\eta) \backslash B_r(X_\eta).$$

It satisfies

$$\begin{cases} \triangle \phi_\eta = 0 & \text{in } B_R(X_\eta) \backslash B_r(X_\eta), \\ \phi_\eta = 1 & \text{on } \partial B_R(X_\eta), \\ \phi_\eta = 0 & \text{on } \partial B_r(X_\eta). \end{cases}$$

By the maximum principle, $u > \phi_\eta$ in $\Omega \cap (B_R(X_\eta) \backslash B_r(X_\eta))$, and on $(\partial B_r) \cap \Gamma_\eta \cap \{x > x_0\}$,

$$c_\eta = |\nabla u| \geq |\nabla \phi_\eta| = \frac{1}{r} \frac{1}{\log \frac{R}{r}} \quad \text{on } (\partial B_r) \cap \Gamma_\eta \cap \{x > x_0\},$$

and proof for the lemma is complete.

Using Lemma 5.5 and the fact that $f'_\eta(s)$ is a step function with jump at $s = \lambda$, we can show that $\mathrm{vol}(\{u = 0\}) = \lambda$ if η is sufficiently small. In fact (see [7]), if $\mathrm{vol}(\{u = 0\}) < \lambda$, then we make a variation of $\{u = 0\}$ with small volume change $\delta v > 0$ by perturbating $\Gamma_\eta \cap \{x > x_0\}$ toward the region $\{u > 0\}$. Such a variation will induce the change $\delta f_\eta = \frac{1}{\eta} \delta v$ and $\delta K = c_\eta^2 \delta v + o(\delta v) \leq C \delta v$ (by Lemma 5.5), where K is the functional induced by the Dirichlet integral. Then

$$\delta J_\eta = \delta K - \delta f_\eta \leq C \delta v - \frac{1}{\eta} \delta v < 0 \qquad \text{if } \eta < \frac{1}{C},$$

a contradiction to u being a minimizer. Similarly, we can derive a contradiction if $\mathrm{vol}(\{u = 0\}) > \lambda$. Therefore, $u_\lambda = u$ is a solution of the problem (5.3). By (5.10) and Lemmas 5.3–5.4, u_λ possesses the following properties:

(5.16)

(1) $u_\lambda \in C^{0,1}(\Omega)$, $\triangle u_\lambda = 0$ in $\{u_\lambda > 0\}$;

(2) $\Gamma_\lambda = \Omega \cap \partial\{u_\lambda > 0\} \cap \{x > x_0\}$ is analytic and there exists a continuous function f such that
$D_\lambda = \{u_\lambda = 0\} = \{(x, y); \quad -f(x) \leq y \leq f(x), \quad x \in \mathcal{I}\}$;

(3) $|\nabla u_\lambda| = c_\lambda$ (c_λ is an unprescribed constant) on Γ_λ.

LEMMA 5.6. *For* $0 < \lambda < \mathrm{vol}(B_\epsilon)$,

$$\int_\Omega |\nabla u_\lambda|^2 \le 4\pi \left\{ \left| \log \frac{\lambda}{\epsilon^2 \pi} \right| \right\}^{-1}.$$

Proof. By the assumption on λ, there exists a concentric disc $B_\mu \subset B_\epsilon$ with radius $\mu = \left[\frac{\lambda}{\pi}\right]^{1/2}$. Introduce the function

$$w = \begin{cases} 1 & \text{in } \Omega \backslash B_\epsilon, \\ 1 - \frac{\log \frac{|X - X_\epsilon|}{\delta}}{\log \frac{\mu}{\epsilon}} & \text{in } B_\epsilon \backslash B_\mu, \\ 0 & \text{in } B_\mu, \end{cases}$$

where X_ϵ is the center of B_ϵ. Then $\mathrm{vol}(\{w = 0\}) = \lambda$, so that $w \in K_{\phi,\lambda}$ and

$$\int_\Omega |\nabla u_\lambda|^2 \le \int_\Omega |\nabla w|^2 = 4\pi \left\{ \left| \log \frac{\lambda}{\epsilon^2 \pi} \right| \right\}^{-1}.$$

Let $D_\lambda = \{x \in \Omega; \quad u_\lambda = 0\}$. We shall show that ∂D_λ does not touch the obstacle if δ is sufficiently small. For this purpose, we first estimate c_λ from above.

LEMMA 5.7. *Suppose* $0 < \lambda < \mathrm{vol}(B_\epsilon)$, $0 < \delta < \sqrt{\pi\lambda}$. *Then*

$$c_\lambda \le 4\sqrt{\frac{\pi}{\lambda}} \left\{ \left| \log \frac{\lambda}{\epsilon^2 \pi} \right| \right\}^{-1}.$$

Proof. By Lemma 5.6 and (5.16),

$$\begin{aligned} 4\pi \left\{ \left| \log \frac{\lambda}{\epsilon^2 \pi} \right| \right\}^{-1} &\ge \int_\Omega |\nabla u_\lambda|^2 \\ &= \int_{\partial\Omega} \frac{\partial u_\lambda}{\partial n} = -\int_{\partial D_\lambda} \frac{\partial u_\lambda}{\partial n} \\ &\ge c_\lambda (\partial D_\lambda) \cap \{x > x_0\} = c_\lambda l_{\Gamma_\lambda}. \end{aligned} \tag{5.17}$$

Using the isoperimetric inequality (2.12), we have

$$\begin{aligned} l_{\Gamma_\lambda} &= l_{\partial D_\lambda} - l_{(\partial D_\lambda) \cap \{x = x_0\}} \\ &\ge \sqrt{4\pi \mathrm{vol}(D_\lambda)} - \delta = 2\sqrt{\pi\lambda} - \delta \ge \sqrt{\pi\lambda} \qquad \text{if} \quad \delta < \sqrt{\pi\lambda}. \end{aligned} \tag{5.18}$$

Combining (5.17) and (5.18), the lemma follows.

LEMMA 5.8. $D_\lambda \cap \{x = x_0\} = \emptyset$ *for all small* δ.

Proof. Suppose the assertion is not true. Then, for $\delta = \delta_j \to 0$,

$$D_\lambda \cap \{x = x_0\} \ne \emptyset.$$

We shall derive a lower bound for c_λ. Since D_λ does not contain an isolated point, we choose $x_\delta > x_0$ close to x_0 such that $D_\lambda \cap \{x = x_\delta\} \ne \emptyset$. Take a disc $B_{2\delta}(X_\delta)$ with $X_\delta = (x_\delta, y_\delta)$ such that $B_{2\delta}$ lies above ∂D_λ, $\partial B_{2\delta} \cap \partial D_\lambda \ne \emptyset$. Since $y_\delta \ge 2\delta$, the concentric disc B_δ lies outside Ω (see Fig. 5.1).

Introduce the function

$$\psi(X) = 1 - \frac{\log \frac{|X - X_\delta|}{\delta}}{\log 2}.$$

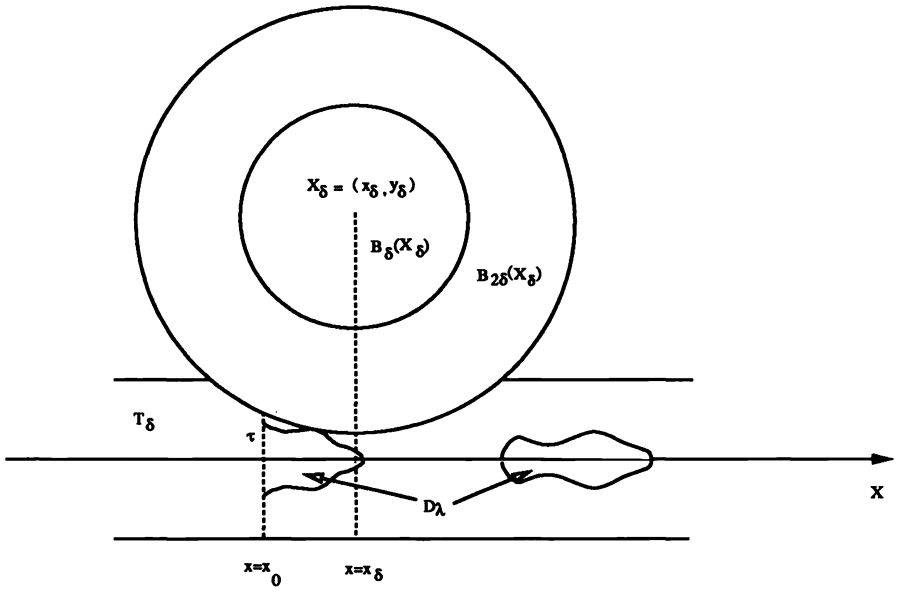


FIG. 5.1

It is the solution of the problem

$$\begin{cases} \triangle\psi = 0 & \text{in } B_{2\delta}\backslash B_\delta, \\ \psi = 1 & \text{on } \partial B_\delta, \\ \psi = 0 & \text{on } \partial B_{2\delta}. \end{cases}$$

By the maximum principle, $\psi < u_\lambda$ in $(B_{2\delta}\backslash B_\delta) \cap \Omega$. We shall use the maximum principle again to derive the lower bound of c_λ. Since the condition $|\nabla u_\lambda| = c_\lambda$ may not hold on $\partial D_\lambda \cap \{x = x_0\}$, we consider the cases

$$\text{(a)} \quad \partial B_{2\delta} \cap \partial D_\lambda \cap \{x > x_0\} \neq \emptyset$$

and

$$\text{(b)} \quad \partial B_{2\delta} \cap \partial D_\lambda \cap \{x = x_0\} \neq \emptyset$$

separately.

If case (a) occurs, then on $\partial B_{2\delta} \cap \partial D_\lambda \cap \{x > x_0\}$,

$$\frac{1}{2\delta}\frac{1}{\log 2} = |\nabla\psi| \leq |\nabla u_\lambda| = c_\lambda. \tag{5.19}$$

It remains to consider case (b). Since $B_{2\delta}$ is centered at $X_\delta = (x_\delta, y_\delta)$ with $x_\delta > x_0$, case (b) implies that $\partial D_\lambda \cap \{x = x_0\}$ contains a line segment τ (see Fig. 5.1). Near the point $X_* \in \partial B_{2\rho} \cap \partial D_\lambda \cap \{x = x_0\}$, the boundary of the domain $\Omega\backslash D_\lambda$ consists of the segment τ and the curve Γ_λ, which is the graph of the continuous function $y = f(x)$. Along τ, $\partial u_\lambda/\partial y = 0$ and Γ_λ, $\limsup_{X\to X_*} \partial u_\lambda/\partial y \leq c_\lambda$. Since $\partial u_\lambda/\partial y$ is bounded near X_*, by Lemma 6.7 in Chapter 2 of [15] (Phragmén–Lindelöf-type lemma), we have

$$\limsup_{X\in\Omega\backslash D_\lambda, X\to X_*} \frac{\partial u_\lambda}{\partial y} \leq c_\lambda. \tag{5.20}$$

Also, since

$$\psi < u_\lambda \quad \text{near} \quad X_* \quad \text{and} \quad \phi = u_\lambda \quad \text{at} \quad X_*,$$

we have

$$\frac{1}{2\delta}\frac{1}{\log 2}\frac{y_\delta - y_*}{|X_* - X_\delta|} = \frac{\partial \phi}{\partial y}(X_*) \le \limsup_{X \in \Omega \backslash D_\lambda, X \to X_*} \frac{\partial u_\lambda}{\partial y} \le c_\lambda. \tag{5.21}$$

Choosing x_δ so close to x_0 that $(y_\delta - y_*)/|X_* - X_\delta| \ge \frac{1}{2}$ (see Fig. 5.1), we have

$$c_\lambda \ge \frac{1}{4\delta \log 2}. \tag{5.22}$$

Combining (5.19) and (5.22), we have established a lower bound $\frac{1}{4\delta \log 2}$ for c_λ. However, this lower bound contradicts Lemma 5.7 as $\delta = \delta_j \to 0$. The proof for Lemma 5.8 is complete.

Proof of Theorem 5.1. By (5.16) and Lemma 5.8, u_λ satisfies following equations:

$$\begin{cases} \triangle u_\lambda = 0 & \text{in} \quad \Omega \backslash D_\lambda, \\ |\nabla u_\lambda| = c_\lambda & \text{on} \quad \Gamma_\lambda = \partial D_\lambda. \end{cases}$$

Note that $\min_{K_{\phi,\lambda}} J(v)$ is continous with respect to λ and δ (see the proof for Lemma 2.3); proceeding as in the proof for Theorem 2.1, there exist λ_I, δ_I such that $c_{\lambda_I} = I/l_{\Gamma_{\lambda_I}}$. Therefore $(u_{\lambda_I}, \Gamma_{\lambda_I})$ is a nonsymmetric solution of (2.1).

Next we construct a symmetric solution, and this will establish nonuniqueness.

Introduce

$$H_s^1(\Omega) = \{v \in H^1(\Omega); \quad v(x,y) = v(-x,y), \quad v(x,y) = v(x,-y)\}$$

and

$$K_{s,\lambda} = K_\lambda \cap H_s^1(\Omega),$$

where K_λ is given by (1.5). Then the argument in [7] shows that there exists a function $u_\lambda \in K_{s,\lambda}$, which minimizes

$$J(v) = \int_\Omega |\nabla v|^2 \qquad \text{for} \quad v \in K_{s,\lambda},$$

and u_λ is Lipschitz in Ω satisfying

$$\begin{cases} \triangle u_\lambda = 0 & \text{in} \quad \Omega \backslash D_\lambda, \\ |\nabla u_\lambda| = c_\lambda & \text{on} \quad \partial D_\lambda = \Gamma_\lambda, \end{cases}$$

where Γ_λ is an analytic curve.

By the argument for proving Theorem 2.1, we can show that there exists a $0 < \lambda < \text{vol}(\Omega)$ such that $c_\lambda = I/l_{\Gamma_\lambda}$. Hence (2.1) also has a symmetric solution.

6. An exterior problem. Given a bounded domain $\Omega \subset \mathbb{R}^2$ and a positive parameter I, consider the following problem: find a curve $\gamma \subset \mathbb{R}^2 \backslash \bar{\Omega}$ and a function u such that

$$\begin{cases} \triangle u = 0 & \text{in} \quad \Omega_\gamma, \\ u = 1 & \text{on} \quad \partial\Omega, \\ u = 0 & \text{on} \quad \gamma, \\ |\nabla u| = \dfrac{I}{l_\gamma} & \text{on} \quad \gamma, \end{cases} \tag{6.1}$$

where Ω_γ is the domain bounded by $\partial\Omega$ and γ, and l_γ is the arc length of γ.

THEOREM 6.1. *Suppose $\partial\Omega$ is Lipschitz continuous. Then, for any $I > 0$, problem (6.1) has a solution (γ, u) with γ as an analytic curve. Moreover, if $\partial\Omega$ is star shaped with respect to a point X_0, γ is also star shaped about X_0 and the solution is unique.*

Proof. Introduce the functional

$$J_\lambda(v) = \int_{R^2\setminus\bar{\Omega}} \left(|\nabla v|^2 + \lambda^2 I_{\{v>0\}}\right) dx, \quad \lambda > 0 \tag{6.2}$$

and the admissible class

$$K = \{v \in H^1_{\text{loc}}(R^2); \quad v \equiv 1 \quad \text{on} \quad \bar{\Omega}, \quad v \geq 0 \quad \text{in} \quad \mathbb{R}^2\}. \tag{6.3}$$

Consider the following problem: find $u = u_\lambda \in K$ such that

$$J_\lambda(u) = \min_{v\in K} J_\lambda(v). \tag{6.4}$$

It is known (see [15, Chap. 3]) that (6.4) has a solution u_λ with a bounded support such that

$$\begin{aligned} &(1) \quad u_\lambda \in C^{0,1}(\mathbb{R}^2),\ \Gamma_\lambda = \partial\{u_\lambda > 0\} \text{ is an analytic curve;} \\ &(2) \quad \triangle u_\lambda = 0 \text{ in } \{u_\lambda > 0\} \cap (\mathbb{R}^2\setminus\bar{\Omega}); \\ &(3) \quad |\nabla u_\lambda| = \lambda \text{ on } \Gamma_\lambda. \end{aligned} \tag{6.5}$$

If $\partial\Omega$ is star shaped with respect to X_0, Tepper in [23] showed that Γ_λ is also star shaped with respect to X_0.

Set

$$m(\lambda) = \min_{v\in K} J_\lambda(v). \tag{6.6}$$

Proceeding as in §2, we can prove that

$$\begin{aligned} &(1) \quad m(\lambda) = \lambda l_{\Gamma_\lambda}; \\ &(2) \quad m(\lambda) \text{ is continuous in } (0,\infty). \end{aligned} \tag{6.7}$$

We shall prove that the range of $m(\lambda)$ is $(0,\infty)$.

By the isoperimetric inequality (2.12),

$$l_{\Gamma_\lambda} \geq \sqrt{4\pi \text{vol}(\{u_\lambda \neq 0\})} \geq \sqrt{4\pi \text{vol}(\Omega)}.$$

By (1) in (6.7) it follows that

$$m(\lambda) \to \infty \qquad \text{as} \quad \lambda \to \infty.$$

To evaluate $\lim_{\lambda\to 0} m(\lambda)$, choose a disc $B_R(\tilde{X})$ containing Ω. For any $\mu > 0$, consider the function

$$\phi_\mu(X) = \begin{cases} 1 & \text{for} \quad X \in B_R(\tilde{X})\setminus\bar{\Omega}; \\ 1 - \mu \log \frac{|X-\tilde{X}|}{R} & \text{for} \quad R < |X-\tilde{X}| < Re^{\frac{1}{\mu}}; \\ 0 & \text{for} \quad |X-\tilde{X}| > Re^{\frac{1}{\mu}}. \end{cases}$$

Then $\phi_\mu \in K$ and

$$\begin{aligned} m(\lambda) &\le J_\lambda(\phi_\mu) \\ &\le \int_R^{Re^{\frac{1}{\mu}}} \mu^2 \frac{1}{r^2} r dr d\theta + \lambda^2 \mathrm{vol}(B_{Re^{\frac{1}{\mu}}}(\tilde{X})) \\ &= 2\pi\mu + \pi R^2 (\lambda e^{\frac{1}{\mu}})^2. \end{aligned}$$

Since the function

$$f(\mu) = \mu e^{-\frac{1}{\mu}}$$

is monotonically increasing in $(0, \infty)$ and

$$\lim_{\mu \to 0+} f(\mu) = 0, \quad \lim_{\mu \to \infty} f(\mu) = \infty,$$

we introduce its inverse function $\mu(\lambda) = f^{-1}(\lambda)$. Clearly, $\mu(0+) = 0$ and $\lambda e^{1/\mu(\lambda)} = \mu(\lambda)$.

Now replace ϕ_μ in J_λ by $\phi_{\mu(\lambda)}$. Then

$$m(\lambda) \le 2\pi\mu(\lambda) + \pi R^2 \mu(\lambda)^2.$$

Letting $\lambda \to 0$ gives $m(\lambda) \to 0$.

Hence, for any $I > 0$, there exists a $\lambda_I > 0$ such that $m(\lambda_I) = I$. By (3) in (6.5) and (1) in (6.7), $(u_{\lambda_I}, \Gamma_{\lambda_I})$ is a solution for (6.1) with $\gamma = \Gamma_{\lambda_I}$. This accomplishes existence.

Suppose that $\partial\Omega$ is star shaped with respect to $X_0 \in \Omega$. We shall apply the well-known Lavrentiev principle [18] to prove uniqueness for problem (6.1).

Denote by G_γ the bounded domain enclosed by γ. For any $0 < \rho < 1$, consider the similarity transformation $X \to \rho(X - X_0)$ which was used in [15] and [23] for jet and cavity problems. Define

$$\begin{aligned} G_\gamma^\rho &= \{\rho(X - X_0); \quad X \in G_\gamma\}, \\ u^\rho(X) &= u\left(\frac{X}{\rho} + X_0\right) \qquad \text{in} \quad G_\gamma^\rho. \end{aligned}$$

To prove uniqueness, suppose there are two solutions (u_i, γ_i) $(i = 1, 2)$ of (6.1). We claim that neither $G_{\gamma_1} \subset G_{\gamma_2}$ nor $G_{\gamma_2} \subset G_{\gamma_1}$. Otherwise, if $G_{\gamma_1} \subset G_{\gamma_2}$ then, by the maximum principle,

$$u_1 < u_2 \qquad \text{in} \quad G_{\gamma_1} \backslash \bar{\Omega}$$

and on $\partial\Omega$, where $u_1 \equiv u_2 \equiv 1$,

$$|\nabla u_1| > |\nabla u_2|,$$

contradicting

$$I = \int_{\partial\Omega} |\nabla u_1| ds = \int_{\partial\Omega} |\nabla u_2| ds.$$

Hence γ_1 and γ_2 must intersect. It follows that there exists $0 < \rho_1 < 1$ such that

$$G_{\gamma_1}^{\rho_1} \subset G_{\gamma_2}, \qquad \partial G_{\gamma_1}^{\rho_1} \cap \partial G_{\gamma_2} \neq \emptyset.$$

Note that $u_1^{\rho_1} \le u_2$ on $\partial(G_{\gamma_1}^{\rho_1} \backslash \bar{\Omega})$. By the maximum principle, $u_1^{\rho_1} < u_2$ in $G_{\gamma_1}^{\rho_1} \backslash \bar{\Omega}$. Since $u_1^{\rho_1} = u_2 = 0$ on $\partial G_{\gamma_1}^{\rho_1} \cap \partial G_{\gamma_2}$, it follows by the maximum principle that

$$\frac{1}{\rho_1} \frac{I}{l_{\gamma_1}} = |\nabla u_1^{\rho_1}| < |\nabla u_2| = \frac{I}{l_{\gamma_2}} \qquad \text{on} \quad \partial G_{\gamma_1}^{\rho_1} \cap \partial G_{\gamma_2},$$

i.e.,

$$l_{\gamma_2} < \rho_1 l_{\gamma_1}. \tag{6.8}$$

By symmetry, there exists $0 < \rho_2 < 1$ such that

$$l_{\gamma_1} < \rho_2 l_{\gamma_2}. \tag{6.9}$$

Combining (6.8) and (6.9), we get

$$l_{\gamma_2} < \rho_1 \rho_2 l_{\gamma_2} < l_{\gamma_2}$$

a contradiction.

Acknowledgments. The author thanks his adviser, Professor Avner Friedman, for his guidence and valuable suggestions. He is grateful for the referee's comments, which allowed him to improve his manuscript.

REFERENCES

[1] A. ACKER, *Heat flow inequalities with application to heat flow optimization*, SIAM J. Math. Anal., 8 (1977), pp. 604–618..

[2] ———, *A free boundary optimization problem*, SIAM J. Math. Anal., 9 (1978), pp. 1179–1191.

[3] ———, *Interior free boundary problems for the Laplace equation*, Arch. Rational Mech. Anal., 75 (1981), pp. 157–168.

[4] ———, *On the qualitative theory of parametrized families of free boundaries*, J. Reine Angew. Math., 393 (1989), pp. 134–169.

[5] ———, *The Bernoulli free-bounary problem—uniqueness and elliptic ordering of solutions*, Sympos. Math., 30 (1989), pp. 1–15.

[6] ———, *Uniqueness and monotonicity of solutions for the interior Bernoulli free boundary problem in the convex, n-dimensional case*, Nonlinear Anal., 12 (1989), pp. 1409–1425.

[7] N. AGUILERA, H. W. ALT, AND L. A. CAFFARELLI, *An optimization problem with volume constraint*, SIAM J. Control Optim., 24 (1986), pp. 191–198.

[8] H. W. ALT AND L. A. CAFFARELLI, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 325 (1981), pp. 105–144.

[9] H. W. ALT, L. A. CAFFARELLI, AND A. FRIEDMAN, *Axially symmetric jet flows*, Arch. Rational Mech. Anal., 81 (1983), pp. 97–149.

[10] A. BADZHADI AND A. S. DEMIDOV, *Existence, nonexistence and regularity theorems in a problem with a free boundary*, Math. USSR-Sb., 50 (1985), pp. 67–84.

[11] A. BEURLING, *On free-boundary problems for the Laplace equation*, Sem. on Analytic Functions 1, Inst. for Advanced Study, Princeton, NJ, 1957, pp. 248–263.

[12] L. A. CAFFARELLI AND J. SPRUCK, *Convexity properties of solutions to some classical variational problems*, Comm. Partial Differential Equations, 7 (1982), pp. 1337–1379.

[13] A. S. DEMIDOV, *The form of a steady plasma subject to the skin effect in a tokamak with non-circular cross section*, Nuclear Fusion, 15 (1975), pp. 765–768.

[14] ———, *Equilibrium form of a steady plasma*, Phys. Fluids, 21 (1978), pp. 902–904.

[15] A. FRIEDMAN, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[16] B. GIDAS, W-N. NI, AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[17] H. KAWARADA, T. SAWAGURI, AND H. IMAI, *An approximate resolution of a free boundary problem appearing in the equilibrium plasma by means of conformal mapping*, Japan J. Appl. Math., 6 (1989), pp. 331–340.

[18] M. A. LAVRENTIEV AND B. W. SCHABAT, *Methoden der Komplexen Funktionentheorie*, VEB Deutscher Verlag der Wissenshaften, Berlin, 1967.

[19] C. B. MORREY, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.

[20] M. NATORI AND H. KAWARADA, *An application of the integrated penalty method to boundary problem of Laplace equation*, Numer. Funct. Anal. Optim., 3 (1981), pp. 1–17.

[21] G. POLYA AND G. SZEGÖ, *Isoperimetric Inequalities in Mathematics Physics*, Princeton University Press, Princeton, NJ, 1951.

[22] J. SERRIN, *A symmetry problem in potential theory*, Arch. Rational Mech. Anal., 43 (1971), pp. 304–318.

[23] D. F. TEPPER, *On a free-boundary problem, the starlike case*, SIAM J. Math. Anal., 6 (1975), pp. 503–505.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1184-1198, September 1995

006

LOWER SEMICONTINUITY CONDITIONS FOR FUNCTIONALS ON JUMPS AND CREASES*

ANDREA BRAIDES†

Abstract. A class of functionals of the form

$$\mathcal{F}(u) = \int_I |u''(t)|^2\,dt + \int_I |u-g|^2\,dt + \sum_{t\in S} \varphi(t, u(t_-), u(t_+), u'(t_-), u'(t_+)),$$

defined on piecewise H^2 functions is studied, where S is the union of the set of points of discontinuity for u and the set of points of discontinuity for u'; $u(t_-), u(t_+)$ (respectively, $u'(t_-), u'(t_+)$) denote the left-hand and right-hand limits of the function u (respectively, of its derivative u') at the point t. Our main results are Theorems 3.1 and 3.4, where we give necessary and sufficient conditions for the lower semicontinuity of $\mathcal{F}$ in the $L^1(I)$ topology in the case of continuous φ. These conditions are of two types: subadditivity properties and a compatibility condition that takes into account the possibility of approximating a "jump" with jumps and "creases" of increasing slope. We show with some examples that both of these conditions are not necessary when φ is only lower semicontinuous. The results are obtained by using recent techniques introduced by De Giorgi and Ambrosio for functions of bounded variation and adapted by Coscia to piecewise H^2 functions. Existence for variational problems related to a model in image segmentation proposed by Blake and Zisserman is derived.

Key words. lower semicontinuous functionals, free discontinuity problems, image segmentation, functions of bounded variation

AMS subject classifications. 49J05, 49J45, 49Q10, 26A45

1. Introduction.

The fundamental problem in pattern recognition is the deduction of the relevant contour of one or more objects from an "input" picture. In a variational approach proposed by Mumford and Shah [12], this contour is modeled as the closed set K such that for some function $u \in C^1(\Omega \setminus K)$ the pair (u, K) attains the minimum value of the integral

$$\int_{\Omega\setminus K} |\nabla u(x)|^2\,dx + \mathcal{H}^1(K) + \lambda \int_\Omega |u(x) - g(x)|^2\,dx. \tag{1.1}$$

The function g represents the original picture, defined on a bounded domain $\Omega \subset \mathbf{R}^2$; $\lambda > 0$ is a suitable constant and $\mathcal{H}^1$ is the one-dimensional Hausdorff measure. The function u represents a good (discontinuous) approximation of the datum g.

This model presents some drawbacks when we also want to detect crease discontinuities of the function u: it is easy to see that for some data g the behaviour of the solution u differs greatly from the input. In order to overcome the inaccuracies

*Received by the editors January 6, 1993; accepted for publication (in revised form) January 3, 1994. The problem of the structure of lower semicontinuous functionals on jumps and creases was addressed by De Giorgi at a Centro Internazionale Ricerche Matematiche workshop on calculus of variations and geometric measure theory, held in Trento in March 1992. This work is part of the Consiglio Nazionale delle Ricerche project Irregular Variational Problems—Discontinuous Structures and was entirely carried out while the author was visiting the Scuola Internazionale Studi Superiori Avanzati in Trieste.

†Dipartimento di Elettronica per l'Automazione, Università di Brescia, via Valotti 9, I-25060 Brescia, Italy.

of this model, Blake and Zisserman proposed to modify the functional (1.1) by introducing second-order derivatives, and indeed their numerical computations show better qualitative approximations of the data (see [4]).

While the model of Mumford and Shah can be included in the framework of functionals defined on the space of *special* functions of bounded variation on Ω, introduced by De Giorgi and Ambrosio (see [10], [1]), it is not clear whether an analogous weak formulation is possible that takes into account higher-order derivatives (see the appendix). In dimension one, though, it is possible to consider functionals of the form

$$\mathcal{F}(u) = \int_I |u''(t)|^2\,dt + \int_I |u-g|^2\,dt + \beta J_0(u) + \alpha J_1(u), \tag{1.2}$$

where $J_0(u)$ is the number of *jump points* of u and $J_1(u)$ is the number of *crease points* of u (i.e., jump points for u' which are not jump points for u) on the space $\mathcal{H}^2(I)$ of piecewise H^2 functions. These functionals still model some problems in segmentation theory (see [4], [8]) and represent the Blake and Zisserman one-dimensional version of (1.1).

In a recent paper, Coscia [8] has proven a compactness result in $\mathcal{H}^2(I)$ with respect to the $L^1(I)$ topology on the sublevel sets $\{\mathcal{F}(u) \le c\}$, when both coefficients α and β are strictly positive and $g \in L^2(I)$. Moreover, she has also shown that a necessary and sufficient condition for the $L^1(I)$-lower semicontinuity of $\mathcal{F}$ is that

$$0 < \alpha \le \beta \le 2\alpha, \tag{1.3}$$

obtaining, in this case, the existence of a solution for the corresponding minimum problem.

This work is devoted to the analysis of necessary and sufficient conditions on the function φ that assure the lower semicontinuity in some natural topology of the general functional

$$\mathcal{F}(u) = \int_I |u''(t)|^2\,dt + \int_I |u-g|^2\,dt + \sum_{t\in S} \varphi(t, u(t_-), u(t_+), u'(t_-), u'(t_+)), \tag{1.4}$$

where S is the union of the set of jump points and the set of crease points for u, and $u(t_-), u(t_+)$ (respectively, $u'(t_-), u'(t_+)$) denote the left-hand and right-hand limits of the function u (respectively, of its derivative u') at the point t. Our main results are Theorems 3.1 and 3.4, where we give necessary and sufficient conditions in the case of continuous φ. These conditions are of two types: subadditivity properties, that assure that it is not convenient to "split" a jump or a crease into more jumps and creases, and a compatibility condition that takes into account the possibility of approximating a jump with jumps and creases of increasing slope. It is interesting to confront these conditions with the general lower semicontinuity conditions for functionals with jumps (see the paper by Ambrosio and Braides [3]; see also [5], [6], and [2]), and remark that, since our functional does not control the first derivative, we cannot simply apply those lower semicontinuity conditions to the pair (u, u'). For a general introduction to lower semicontinuity problems in the calculus of variations, we refer to [7] and [9].

The plan of the paper is as follows. Section 2 is devoted to preliminaries; in particular we recall the compactness result by Coscia and prove the simple Lemma 2.3 that provides some links between first and second derivatives for sequences with bounded energy for the functional $\mathcal{F}$. In §3 we state and prove the main result and apply it

to recover the lower semicontinuity theorems of [8]. Section 4 is devoted to the non-continuous case. We first deal with the case of φ invariant by translations and show that necessary conditions are the lower semicontinuity of φ, subadditivity, and some kind of compatibility at infinity. Finally, we show with two counterexamples that, in the general case, these conditions (and hence also the conditions of Theorems 3.1 and 3.4) are not necessary.

2. Notation and preliminaries. We use standard notation for Lebesgue and Sobolev spaces; in particular we make use of the spaces $\mathrm{L}^1(I)$, $\mathrm{L}^2(I)$, and $\mathrm{H}^2(I)$, where I is an open bounded interval of $\mathbf{R}$. We denote with $\#$ the *counting measure*, and with $\overline{\mathbf{R}} = \mathbf{R} \cup \{-\infty, +\infty\}$ the *extended real line*.

Following the work of Coscia, we define the space $\mathcal{H}^2(a,b)$ as the space of functions $u \in \mathrm{L}^2(a,b)$ for which it is possible to find a finite number of points $x^0 = a < x^1 < \cdots < x^k < x^{k+1} = b$ such that $u \in \mathrm{H}^2(x^j, x^{j+1})$ for every $j = 0, \dots, k$. Note that if $u \in \mathcal{H}^2(a,b)$, then u and u' are bounded and absolutely continuous on each subinterval (x^j, x^{j+1}). In particular there exist the right-hand and left-hand limits

$$u(x^j{+}),\ u(x^j{-}), \qquad u'(x^j{+}),\ u'(x^j{-}), \tag{2.1}$$

for every $j = 0, \dots, k+1$ (except at $x^0{-}$ and $x^{k+1}{+}$, of course), and they are finite.

We will regard the functions u' and u'' as defined almost everywhere (a.e.) on the whole interval (a,b). Moreover, the functions $x \mapsto u(x_+)$, $x \mapsto u'(x_+)$ are defined everywhere on $[a,b)$, and the functions $x \mapsto u(x_-)$, $x \mapsto u'(x_-)$ are defined everywhere on $(a,b]$.

We define the set S_u of *jump points* of u as the set of those $x \in (a,b)$ for which we have $u(x_+) \neq u(x_-)$. In the same way we define the set $S_{u'}$ of jump points of u' for which $u'(x_+) \neq u'(x_-)$. The set of *crease points* of u is defined as $S_{u'} \setminus S_u$.

Note that we have $S_{u'} \cup S_u \subset \{x^1, \dots, x^k\}$, and that if $x^j \notin (S_{u'} \cup S_u)$ and $j \neq 0$, $j \neq k+1$, then $u \in \mathrm{H}^2(x^{j-1}, x^{j+1})$ and, therefore, the point x^j can be removed from the subdivision. Hence we can and will suppose that

$$\{x^1, \dots, x^k\} = S_{u'} \cup S_u.$$

We say that a sequence $(u_h) \subset \mathcal{H}^2(a,b)$ converges weakly in $\mathcal{H}^2(a,b)$ to $u \in \mathcal{H}^2(a,b)$ if we have $\sup_h \#(S_{u_h} \cup S_{u_h'}) < +\infty$, $u_h \to u$ strongly in $\mathrm{L}^1(a,b)$, $u_h' \to u'$ a.e. in (a,b), and $u_h'' \rightharpoonup u''$ weakly in $\mathrm{L}^2(a,b)$. This is the "right" notion of weak convergence to consider as shown by the following compactness lemma by Coscia [8].

LEMMA 2.1 (compactness, Coscia [8]). *Let us consider a bounded open interval $I \subset \mathbf{R}$ and a sequence of functions $(u_h) \subset \mathcal{H}^2(a,b)$ such that*

(i) *the sequence (u_h) is bounded in $\mathrm{L}^2(I)$;*
(ii) *the sequence (u_h'') is bounded in $\mathrm{L}^2(I)$;*
(iii) *we have $\sup_h \#(S_{u_h'} \cup S_{u_h}) < +\infty$.*

Then there exists a subsequence (u_{h_k}) and a function $u \in \mathcal{H}^2(I)$ such that $u_{h_k} \rightharpoonup u$ weakly in $\mathcal{H}^2(I)$.

We will study the sequential lower semicontinuity of functionals of the form

$$F(u) = \sum_{t \in S_u \cup S_{u'}} \varphi(t, u(t_-), u(t_+), u'(t_-), u'(t_+)) \tag{2.2}$$

with respect to the weak convergence of $\mathcal{H}^2(I)$.

Lower semicontinuity conditions, together with the compactness result, Lemma 2.1, will lead to an existence theorem for a general class of minimum problems. We will make the assumption that $\varphi \geq c > 0$, which is satisfied by the model functional (1.2). This condition is necessary for setting the problem in $\mathcal{H}^2$; moreover, it assures the equiboundedness of the number of crease and jump points on minimizing sequences, and hence the applicability of Lemma 2.1. We refer the interested reader to the appendix for a formulation of the problem under weaker coerciveness assumptions in spaces of special functions of bounded variation.

PROPOSITION 2.2 (an existence result). *Let us consider $g \in \mathrm{L}^2(I)$. Let us suppose that the functional F in (2.2) is sequentially lower semicontinuous with respect to the weak convergence of $\mathcal{H}^2(I)$, and $\varphi \geq c > 0$. Then, there exists a solution $u \in \mathcal{H}^2(I)$ of the segmentation problem*

$$\min\Big\{ \int_I |u''|^2\,dt + \int_I |u-g|^2\,dt + \sum_{t\in S_u\cup S_{u'}} \varphi(t,u(t_-),u(t_+),u'(t_-),u'(t_+)) \ : \ u\in\mathcal{H}^2(I)\Big\}.$$

Proof. We can rewrite our problem as

$$\min\Big\{\int_I |u''|^2\,dt + \int_I |u-g|^2\,dt + F(u) \ : \ u\in\mathcal{H}^2(I)\Big\}, \tag{2.3}$$

where F is given by (2.2). Moreover, since $\varphi \geq c > 0$, we have

$$F(u) \geq c\,\#(S_u \cup S_{u'}). \tag{2.4}$$

Let us consider a minimizing sequence (u_h). By (2.3) and (2.4) this sequence satisfies hypotheses (i)–(iii) of Lemma 2.1. Hence we can suppose (up to subsequences) that $u_h \rightharpoonup u$ weakly in $\mathcal{H}^2(I)$. By the weak convergence in $\mathrm{L}^2(I)$ of u_h'' to u' we get

$$\int_I |u''|^2\,dt \leq \liminf_h \int_I |u_h''|^2\,dt; \tag{2.5}$$

by the convergence of u_h to u in $\mathrm{L}^1(I)$, and the weak lower semicontinuity of the L^2-norm, we get

$$\int_I |u-g|^2\,dt \leq \liminf_h \int_I |u_h-g|^2\,dt.$$

Finally, by hypothesis F is sequentially lower semicontinuous with respect to the weak convergence of $\mathcal{H}^2(I)$, hence $F(u) \leq \liminf_h F(u_h)$. These inequalities show that u attains the minimum in (2.3). □

We conclude this section by proving an elementary lemma that will be needed in what follows. It shows that, even though the functionals considered here do not directly weigh the first derivative, u' is not allowed to have an arbitrary behaviour.

LEMMA 2.3. *Let $t_1, t_2, w_1, w_2, \zeta_1, \zeta_2 \in \mathbf{R}$ with $t_1 < t_2$. Then we have*

$$\min\Big\{\int_{t_1}^{t_2} |u''|^2\,dt \ : u\in \mathrm{H}^2(t_1,t_2),\ u(t_1)=w_1, u(t_2)=w_2, u'(t_1)=\zeta_1, u'(t_2)=\zeta_2\Big\}$$
$$= \frac{1}{t_2-t_1}\Big(3\big((\zeta_1+\zeta_2) - \frac{2}{t_2-t_1}(w_1-w_2)\big)^2 + (\zeta_1-\zeta_2)^2\Big).$$

Proof. It suffices to remark that the Euler–Lagrange equation related to the aforementioned minimum problem

$$u^{(iv)} = 0$$

has a unique solution satisfying the given boundary conditions. Then it is easy to compute this polynomial of degree 3 and check the minimum value. □

Remark 2.4. Let us consider two sequences of points (t_h^1), (t_h^2) with $t_h^1 < t_h^2$, converging to the same point t, and a sequence of functions $(u_h) \subset \mathrm{H}^2(t_h^1, t_h^2)$ such that

$$\sup_h \int_{t_h^1}^{t_h^2} |u_h''|^2 \, dt < +\infty.$$

If we have $u_h'(t_h^1) \to \zeta_1$, and $u_h'(t_h^2) \to \zeta_2$, then $\zeta_1 = \zeta_2$. In fact, from Lemma 2.3 we obtain that

$$\big(u_h'(t_h^1) - u_h'(t_h^2)\big)^2 \le c(t_h^2 - t_h^1) \to 0.$$

Moreover, if $\zeta_1 = \zeta_2 = \zeta \in [-\infty, +\infty]$, then we must have

$$\lim_h \frac{u(t_h^2) - u(t_h^1)}{t_h^2 - t_h^1} = \zeta.$$

In particular, if $\zeta \notin \{-\infty, +\infty\}$ then we must have

$$\lim_h (u(t_h^2) - u(t_h^1)) = 0,$$

and if $u(t_h^1) \to u_1$, $u(t_h^2) \to u_2$ with $u_2 > u_1$ (respectively, $u_2 < u_1$) then $\zeta = +\infty$ (respectively, $-\infty$).

3. A lower semicontinuity theorem. In this section we prove necessary and sufficient conditions for the sequential lower semicontinuity with respect to the weak convergence in $\mathcal{H}^2(I)$ of functionals of the form

$$F(u) = \sum_{t \in S_u \cup S_{u'}} \varphi(t, u(t_-), u(t_+), u'(t_-), u'(t_+)) \tag{3.1}$$

in the case of a *continuous* "integrand" φ. This result will not directly include some important cases considered in the literature, where φ is supposed to be only lower semicontinuous; in any case, we will show in Theorem 3.5 how to treat easily these cases using our Theorem 3.1 and an approximation procedure. Our main result is the following theorem.

THEOREM 3.1 (lower semicontinuity: the case of continuous φ). *Let us consider a continuous function $\varphi : I \times \mathbf{R}^2 \times \overline{\mathbf{R}}^2 \to [0, +\infty]$. Then, necessary and sufficient conditions for the functional F in* (3.1) *to be sequentially lower semicontinuous with respect to the weak convergence of $\mathcal{H}^2(I)$ are as follows:*

(i) *(subadditivity) for every $t \in I$ and $u, v, w, \xi, \eta, \zeta \in \mathbf{R}$ such that $(u, \xi) \neq (v, \eta)$, we have*

$$\varphi(t, u, v, \xi, \eta) \le \varphi(t, u, w, \xi, \zeta) + \varphi(t, w, v, \zeta, \eta); \tag{3.2}$$

(ii) *(compatibility) for every $t \in I$ and $u, v, w, z, \xi, \eta \in \mathbf{R}$ with $(u, \xi) \neq (v, \eta)$ and $w \neq z$, we have*

$$\varphi(t, u, v, \xi, \eta) \le \varphi(t, u, w, \xi, \zeta) + \varphi(t, z, v, \zeta, \eta), \tag{3.3}$$

where $\zeta = +\infty$ *if* $z > w$ *and* $\zeta = -\infty$ *if* $z < w$.

Remark 3.2. (1) It is worth spending a few words on the function φ and conditions (3.2) and (3.3). In the definition of φ we also take into account the values $-\infty$ and $+\infty$. These correspond to "infinite slope" or a "vertical crease." We also have to take into account these degenerate creases, since a "pure jump" (for which we have to consider $\varphi(t, u(t_-), u(t_+), u'(t_-), u'(t_+))$) can be approximated using slopes diverging to $\pm\infty$ (for which we have to consider $\varphi(t, u(t_-), u(t_-), u'(t_-), \pm\infty)$ and $\varphi(t, u(t_+), u(t_+), \pm\infty, u'(t_+))$—note that we make no assumption on the L^2 norm of the derivative u', and hence steep slopes are allowed). Condition (3.3) accounts exactly for this possibility, excluding the convenience of infinite slopes. Note that if $u < v$, then condition (3.3) in particular yields

$$\varphi(t, u, v, \xi, \eta) \leq \varphi(t, u, u, \xi, +\infty) + \varphi(t, v, v, +\infty, \eta)$$

(and an analogous condition holds if $u > v$). This case corresponds to the approximation of a jump with two creases. The inequality in (3.2) instead shows that "splitting" a jump into two jumps cannot lower the value of F.

(2) By the continuity of φ, (3.2) and (3.3) must also hold for $\xi, \eta, \zeta \in \{-\infty, +\infty\}$.

(3) Since we will never consider points of the form (u, u, ξ, ξ), we can require φ to be defined and continuous, and satisfy (3.2) and (3.3) only outside the "diagonal" $\{(t, u, u, \xi, \xi) \ : \ t \in I, u \in \mathbf{R}, \xi \in \overline{\mathbf{R}}\}$.

(4) Functionals of the form

$$F(u) = \sum_{S_u \cup S_{u'}} \psi(t, u(t_+) - u(t_-), u'(t_+) - u'(t_-))$$

can be dealt with in the same way as in Theorem 3.1 (see Theorem 3.4). Note, however, that the function $\varphi(t, u, v, \xi, \eta) = \psi(t, v - u, \eta - \xi)$ will not satisfy the hypotheses of Theorem 3.1, except for trivial cases.

Before proceeding in the proof of Theorem 3.1 we give, in the following proposition an equivalent form for (3.2) and (3.3) that will be useful in what follows.

PROPOSITION 3.3. *Conditions* (3.2) *and* (3.3) *are equivalent to the requirement that we have*

$$\varphi(t, u, v, \xi, \eta) \leq \sum_{i=1}^{m} \varphi(t, u_-^i, u_+^i, \zeta^i, \zeta^{i+1}) \tag{3.4}$$

for every choice of real numbers $u, v, \xi, \eta \in \mathbf{R}$ *and every choice of* (u_-^i), $(u_+^i) \subset \mathbf{R}$, $(\zeta^i) \subset \overline{\mathbf{R}}$, $i = 1, \ldots, m$, $m \in \mathbf{N}$, *satisfying* $\zeta^i = +\infty$ *if* $u_-^i > u_+^{i-1}$ *and* $\zeta^i = -\infty$ *if* $u_-^i < u_+^{i-1}$, $\zeta^1 = \xi$, $\zeta_m = \eta$, $u_-^1 = u$, $u_+^m = v$.

Proof. Clearly (3.5) implies (3.2) and (3.3). Let us prove the converse. If $m = 1$ the equality is trivial. If $m = 2$ we have two possibilities: either we have $u_+^1 = u_-^2$, in which case we obtain (3.2) with $\zeta = \zeta^2 \in \overline{\mathbf{R}}$ and $w = u_+^1 = u_-^2$, or we have $u_+^1 \neq u_-^2$, in which case (3.4) is equal to (3.3). Then the proof can proceed by induction. Let us consider $u, v, \xi, \eta \in \mathbf{R}$ and (u_-^i), $(u_+^i) \subset \mathbf{R}$, $(\zeta^i) \subset \overline{\mathbf{R}}$, $i = 1, \ldots, m$, as above. Let us first suppose $u_+^{m-1} = u_-^m$ and set $\zeta = \zeta^m \in \overline{\mathbf{R}}$ and $w = u_+^{m-1} = u_-^m$. Then we have

$$\begin{aligned}\sum_{i=1}^{m} \varphi(t, u_-^i, u_+^i, \zeta^i, \zeta^{i+1}) &= \sum_{i=1}^{m-1} \varphi(t, u_-^i, u_+^i, \zeta^i, \zeta^{i+1}) + \varphi(t, w, v, \zeta, \eta) \\ &\geq \varphi(t, u, w, \xi, \zeta) + \varphi(t, w, v, \zeta, \eta) \geq \varphi(t, u, v, \xi, \eta).\end{aligned}$$

Note that we have made use of (3.2), Remark 3.2(2) if necessary, and (3.4) for $m-1$, which also holds true for $\zeta \in \overline{\mathbf{R}}$ by the continuity of φ.

If $u_+^{m-1} < u_-^m$, then we must have $\zeta^m = +\infty$. We therefore obtain

$$\sum_{i=1}^{m} \varphi(t, u_-^i, u_+^i, \zeta^i, \zeta^{i+1}) = \sum_{i=1}^{m-1} \varphi(t, u_-^i, u_+^i, \zeta^i, \zeta^{i+1}) + \varphi(t, u_-^m, v, +\infty, \eta)$$
$$\geq \varphi(t, u, u_+^{m-1}, \xi, +\infty) + \varphi(t, u_-^m, v, +\infty, \eta) \geq \varphi(t, u, v, \xi, \eta).$$

Here we have used (3.3) and (3.4) for $m-1$. In the same way, we deal with the case $u_+^{m-1} > u_-^m$ and $\zeta^m = -\infty$. □

Proof of Theorem 3.1. Without loss of generality we suppose that $I = (-1, 1)$.

Part one: Necessity. Let us fix $t \in (-1, 1)$; we can suppose $t = 0$. Let us consider $u, v, \xi, \eta \in \mathbf{R}$ with $(u, \xi) \neq (v, \eta)$ and define the function $u \in \mathcal{H}^2(-1, 1)$ by setting

$$u(x) = \begin{cases} u + \xi x & \text{if } x < 0, \\ v + \eta x & \text{if } x \geq 0. \end{cases} \tag{3.5}$$

We have $S_u \cup S_{u'} = \{0\}$ and $F(u) = \varphi(0, u, v, \xi, \eta)$. Let us define the sequence $(u_h) \subset \mathcal{H}^2(-1, 1)$ by setting

$$u_h(x) = \begin{cases} u + \xi(x + \frac{1}{h}) & \text{if } x \leq -\frac{1}{h}, \\ w + \zeta x & \text{if } -\frac{1}{h} < x < \frac{1}{h}, \\ v + \eta(x - \frac{1}{h}) & \text{if } x \geq \frac{1}{h}. \end{cases} \tag{3.6}$$

We have $u_h \rightharpoonup u$ weakly in $\mathcal{H}^2(-1, 1)$ and $S_{u_h} \cup S_{u_h'} \subset \{-\frac{1}{h}, \frac{1}{h}\}$, so that

$$F(u_h) \leq \varphi\left(-\frac{1}{h}, u, w - \frac{1}{h}\zeta, \xi, \zeta\right) + \varphi\left(\frac{1}{h}, w + \frac{1}{h}\zeta, v, \zeta, \eta\right).$$

By the lower semicontinuity of F we then have

$$\begin{aligned} \varphi(0, u, v, \xi, \eta) = F(u) &\leq \liminf_h F(u_h) \\ &\leq \liminf_h \Big(\varphi\Big(-\frac{1}{h}, u, w - \frac{1}{h}\zeta, \xi, \zeta\Big) + \varphi\Big(\frac{1}{h}, w + \frac{1}{h}\zeta, v, \zeta, \eta\Big)\Big) \\ &= \varphi(0, u, w, \xi, \zeta) + \varphi(0, w, v, \zeta, \eta); \end{aligned}$$

that is, we have equation (3.2). Let us remark that this inequality is trivial if $(u, \xi) = (w, \zeta)$ or $(v, \eta) = (w, \zeta)$, and hence we do not really use the continuity of φ on the "diagonal" (recall Remark 3.2(3)).

The proof of (3.3) follows the same line, using a different approximating sequence for the function u defined in (3.5). Let us choose $w, z \in \mathbf{R}$ with $w \neq z$, and construct the sequence u_h as follows:

$$u_h(x) = \begin{cases} u + \xi(x + \frac{1}{h}) & \text{if } x \leq -\frac{1}{h}, \\ \frac{1}{2}(w + z) + h\frac{1}{2}(z - w)x & \text{if } -\frac{1}{h} < x < \frac{1}{h}, \\ v + \eta(x - \frac{1}{h}) & \text{if } x \geq \frac{1}{h}. \end{cases} \tag{3.7}$$

Again, $u_h \rightharpoonup u$ weakly in $\mathcal{H}^2(-1, 1)$, $S_{u_h} \cup S_{u_h'} = \{-\frac{1}{h}, \frac{1}{h}\}$, and

$$F(u_h) = \varphi\left(-\frac{1}{h}, u, w, \xi, \frac{h}{2}(z - w)\right) + \varphi\left(\frac{1}{h}, z, v, \frac{h}{2}(z - w), \eta\right).$$

By the lower semicontinuity of F and the continuity of φ, we then get

$$\begin{aligned}\varphi(0,u,v,\xi,\eta) &\le \liminf_h \Big(\varphi\Big(-\frac{1}{h},u,w,\xi,\frac{h}{2}(z-w)\Big)+\varphi\Big(\frac{1}{h},z,v,\frac{h}{2}(z-w),\eta\Big)\Big)\\ &= \varphi(0,u,w,\xi,\zeta)+\varphi(0,z,v,\zeta,\eta),\end{aligned}$$

where $\zeta = \lim_{h\to+\infty}\frac{h}{2}(z-w)$, and hence we have obtained (3.3).

Part two: Sufficiency. Let us fix a sequence $(u_h)\subset \mathcal{H}^2(-1,1)$ converging to u weakly in $\mathcal{H}^2(-1,1)$. Then we can suppose (up to subsequences) that it is $\#(S_{u_h}\cup S_{u_h'})\le N<+\infty$. Let us consider a point $t\in S_u\cup S_{u'}$. We can suppose (up to a further subsequence) that there exist exactly m sequences $(x_h^i)\subset S_{u_h}\cup S_{u_h'}$ converging to t with $x_h^1<x_h^2<\cdots<x_h^m$, and for every $i=1,\dots,m$ we have

$$u_h(x_{h\pm}^i)\to u_\pm^i\in\mathbf{R},\qquad u_h'(x_{h\pm}^i)\to\zeta_\pm^i\in\overline{\mathbf{R}}.$$

Note that by Remark 2.4 we have that $\zeta_+^i=\zeta_-^{i+1}$ for $i=1,\dots,m-1$. Let us define $\zeta^i=\zeta_-^i$ for $i=1,\dots,m$ and $\zeta^{m+1}=\zeta_+^m$. Again, by Remark 2.4, we obtain that $u_+^{i-1}=u_-^i$ whenever $\zeta^i\in\mathbf{R}$, and also $\zeta^i=+\infty$ (respectively, $\zeta^i=-\infty$) whenever $u_+^{i-1}<u_-^i$ (respectively, $u_+^{i-1}>u_-^i$). Note also that

$$u_-^1=u(t_-),\quad u_+^m=u(t_+),\quad \zeta_-^1=u'(t_-),\quad \zeta_+^m=u'(t_+).$$

The quantities (u_-^i), (u_+^i), and (ζ^i), $i=1,\dots,m$, satisfy the hypotheses of Proposition 3.3 with $u=u(t_-)$, $v=u(t_+)$, $\xi=u'(t_-)$, and $\eta=u'(t_+)$. Therefore we have

$$\begin{aligned}\varphi(t,u(t_-),u(t_+),u'(t_-),u'(t_+)) &\le \sum_{i=1}^m \varphi(t,u_-^i,u_+^i,\zeta_-^i,\zeta_+^i)\\ &= \lim_h \varphi(x_h^i,u_h(x_{h-}^i),u_h(x_{h+}^i),u_h'(x_{h-}^i),u_h'(x_{h+}^i)).\end{aligned}$$

Repeating this argument for every $t\in S_u\cup S_{u'}$, we obtain

$$F(u)\le\liminf_h F(u_h).$$

Hence the theorem is proved. □

THEOREM 3.4. *Let $\psi: I\times\mathbf{R}\times\overline{\mathbf{R}}\to[0,+\infty]$ be a continuous function. Necessary and sufficient conditions for the functional*

$$F(u)=\sum_{S_u\cup S_{u'}}\psi(t,u(t_+)-u(t_-),u'(t_+)-u'(t_-))$$

to be lower semicontinuous with respect to the weak convergence of $\mathcal{H}^2(I)$ are

(i) *(subadditivity) for every $t\in I$ and every $u,v,\xi,\eta\in\mathbf{R}$,*

$$\psi(t,u+v,\xi+\eta)\le\psi(t,u,\xi)+\psi(t,v,\eta);$$

(ii) *(compatibility) for every $t\in I$,*

$$\sup\psi(t,\cdot,\cdot)\le\inf\psi(t,\cdot,+\infty)+\inf\psi(t,\cdot,-\infty).$$

Proof. The proof of Theorem 3.1 can be followed, taking some extra care in the proof of Proposition 3.3 and the sufficiency part, possibly extracting further subsequences in order to avoid $+\infty-\infty$ indeterminations. □

We want to apply Theorem 3.4 to re-obtain a lower semicontinuity theorem proven by Coscia [8] for the functional defined on $\mathcal{H}^2(0,1)$ by setting

$$\mathcal{F}(u)=\int_0^1|u''|^2\,dt+\int_0^1|u-g|^2\,dt+\alpha\#(S_{u'}\setminus S_u)+\beta\#(S_u). \tag{3.8}$$

This functional can be written as

$$\mathcal{F}(u)=\int_0^1|u''|^2\,dt+\int_0^1|u-g|^2\,dt+F(u),$$

where F is defined as in Theorem 3.4, and ψ is defined by setting

$$\psi(t,s,\xi)=\psi(s)=\begin{cases}\beta & \text{if } s\neq 0,\\ \alpha & \text{if } s=0.\end{cases}$$

THEOREM 3.5 (A. Coscia [8]). *Let $\alpha,\beta\in\mathbf{R}$ satisfy*

$$0<\alpha\le\beta\le 2\alpha. \tag{3.9}$$

Then the functional $\mathcal{F}$, extended to $+\infty$ on $\mathrm{L}^1(0,1)\setminus\mathcal{H}^2(0,1)$, is lower semicontinuous with respect to the $\mathrm{L}^1(0,1)$ convergence.

Proof. By Lemma 2.1 it suffices to prove that F is lower semicontinuous on $\mathcal{H}^2(0,1)$ with respect to the weak convergence of $\mathcal{H}^2(0,1)$. We cannot directly apply Theorem 3.4 since the function ψ is not continuous, but only lower semicontinuous. Nevertheless, we can write F as the pointwise supremum of a family of functionals F_h for which our result applies. Let us define, for every $h\in\mathbf{N}$,

$$\psi_h(t,s,\xi)=\big(\alpha+h|s|\big)\wedge\beta.$$

Each ψ_h is continuous on $(0,1)\times\mathbf{R}\times\overline{\mathbf{R}}$. Let us check conditions (i) and (ii) of Theorem 3.4. We have

$$\begin{aligned}\psi_h(t,u+v,\xi+\eta)&=(\alpha+h|u+v|)\wedge\beta\le(\alpha+h|u|+h|v|)\wedge\beta\\&\le\big((\alpha+h|u|)\wedge\beta\big)+\big((\alpha+h|v|)\wedge\beta\big)=\psi_h(t,u,\xi)+\psi_h(t,v,\eta)\end{aligned}$$

such that (i) is satisfied. Moreover, we get

$$\psi_h(t,s,\xi)=(\alpha+h|s|)\wedge\beta\le\beta\le 2\alpha=\inf\psi_h(t,\cdot,+\infty)+\inf\psi_h(t,\cdot,-\infty),$$

and hence we also obtain (ii). If we define, for $u\in\mathcal{H}^2(0,1)$,

$$F_h(u)=\sum_{t\in S_u\cup S_{u'}}\psi_h(t,u(t_+)-u(t_-),u'(t_+)-u'(t_-)),$$

then F_h is lower semicontinuous in $\mathcal{H}^2(0,1)$ by Theorem 3.4. The lower semicontinuity of F in $\mathcal{H}^2(0,1)$ follows by observing that $F(u)=\sup_h F_h(u)$. Eventually, the lower semicontinuity of $\mathcal{F}$ can be proven as in Proposition 2.2 using Lemma 2.1. □

Remark 3.6. The proof of Theorem 3.5 can be easily extended to prove the lower semicontinuity in $\mathcal{H}^2(0,1)$ of functionals of the form

$$F(u) = \sum_{t \in S_{u'} \setminus S_u} \psi(u'(t_+) - u'(t_-)) + \beta \#(S_u)$$

with $\psi : \overline{\mathbf{R}} \to [0, +\infty]$ subadditive and continuous, under the hypothesis

$$\sup \psi \leq \beta \leq \psi(-\infty) + \psi(+\infty).$$

This condition is also necessary (see Remark 4.2(2)).

4. Some counterexamples. In the previous section we proved the lower semicontinuity theorem under the hypothesis of continuity of the function φ. This section is devoted to the analysis of the functionals of the form (3.1) in the case of lower semicontinuous φ. We will see, with the aid of some examples, that it is not possible in general to give a simple criterion of lower semicontinuity for these functionals.

Let us first give some necessary conditions for the special class of functionals F, which are even and invariant under the addition of C^1 functions; i.e., $F(-u) = F(u) = F(u + \phi)$ for all $\phi \in C^1$. These functionals depend only on the absolute size of jumps and creases.

PROPOSITION 4.1. *Let $\varphi : [0, +\infty[\times[0, +\infty[\to [0, +\infty]$ be a Borel function, and let us suppose that the functional F defined on $\mathcal{H}^2(I)$ by*

$$F(u) = \sum_{t \in S_u \cup S_{u'}} \varphi\big(|u(t_+) - u(t_-)|, |u'(t_+) - u'(t_-)|\big)$$

is sequentially lower semicontinuous in $\mathcal{H}^2(I)$. Then we must have

(i) *φ is lower semicontinuous on $([0, +\infty[\times[0, +\infty[) \setminus \{(0,0)\}$;*

(ii) *(subadditivity) for every $u, v, \xi, \eta, \in \mathbf{R}$ with $u \neq v$, $\xi \neq \eta$,*

$$\varphi(|u - v|, |\xi - \eta|) \leq \varphi(|u|, |\xi|) + \varphi(|v|, |\eta|);$$

(iii) *(compatibility) for every $u, s \geq 0$ with $(u, s) \neq (0,0)$,*

$$\varphi(u, s) \leq \liminf_{t \to +\infty} (\varphi(0, t) + \varphi(0, t + s)).$$

Proof. (i) We can suppose $0 \in I$. Let us fix $(u, \xi) \neq (0,0)$ and $(u_h, \xi_h) \to (u, \xi)$ with $u, u_h, \xi, \xi_h \geq 0$. Let us consider the functions

$$u_h(t) = \begin{cases} 0 & \text{if } t < 0, \\ u_h + \xi_h t & \text{if } t \geq 0, \end{cases}$$

and

$$u(t) = \begin{cases} 0 & \text{if } t < 0, \\ u + \xi t & \text{if } t \geq 0. \end{cases}$$

Then we obtain

$$\varphi(u, \xi) = F(u) \leq \liminf_h F(u_h) = \varphi(u_h, \xi_h).$$

(ii) Let us define the functions u as in (3.5) and u_h by setting

$$u_h(t) = \begin{cases} u + \xi(t + \frac{1}{h}) & \text{if } t \le -\frac{1}{h}, \\ 0 & \text{if } -\frac{1}{h} < t < \frac{1}{h}, \\ v + \eta(t - \frac{1}{h}) & \text{if } t \ge \frac{1}{h}. \end{cases}$$

It is easy to check that $u_h \to u$ in $\mathcal{H}^2(I)$. Then by the lower semicontinuity of F we get

$$\varphi(|u - v|, |\xi - \eta|) = F(u) \le \liminf_h F(u_h) = \varphi(|u|, |\xi|) + \varphi(|v|, |\eta|).$$

(iii) It suffices to consider the sequence of functions

$$u_h(t) = \begin{cases} 0 & \text{if } t < 0, \\ ht & \text{if } 0 \le t \le \frac{u}{h}, \\ u + \xi(t - \frac{u}{h}) & \text{if } t > \frac{u}{h} \end{cases}$$

that converge in $\mathcal{H}^2(I)$ to the function u in the proof of (i), and apply the lower semicontinuity. □

Remark 4.2. (1) Other necessary conditions different from (iii) can be obtained by following the arguments used in the proof of the necessity of (3.3).

(2) From Proposition 4.1 we obtain that necessary conditions for the lower semicontinuity of the functional $\mathcal{F}$ in (3.8) are $\alpha \le \beta$ (for the lower semicontinuity of φ) and $\beta \le 2\alpha$ (for the compatibility condition (iii)). Hence, recalling Theorem 3.5, we have that condition (3.9) is necessary and sufficient for the lower semicontinuity of $\mathcal{F}$.

If we do not require the continuity of φ at infinity, condition (ii) of Theorem 3.1 (or condition (ii) of Theorem 3.4) is not necessary for semicontinuity as shown by the following example.

Example 4.3. Let us consider the functions $f_1, f_2, f : \mathbf{R} \to \{1, 2\}$, defined by

$$f_1(t) = \begin{cases} 1 & \text{if } t \in 2\mathbf{Z}, \\ 2 & \text{otherwise,} \end{cases} \qquad f_2(t) = \begin{cases} 1 & \text{if } t + 1 \in 2\mathbf{Z}, \\ 2 & \text{otherwise,} \end{cases}$$

and $f = f_1 \wedge f_2$, and let us define $\varphi : \mathbf{R}^4 \to \{1, 2, 3\}$ by setting

$$\varphi(u, v, \xi, \eta) = \begin{cases} f_1(\xi - \eta) & \text{if } u = v > 0, \\ f_2(\xi - \eta) & \text{if } u = v < 0, \\ f(\xi - \eta) & \text{if } u = v = 0, \\ 3 & \text{if } u \ne v,\ uv < 0, \\ 2 & \text{if } u \ne v,\ uv \ge 0. \end{cases}$$

The function φ is lower semicontinuous and we can extend it at infinity by lower semicontinuity. For example, we set

$$\varphi(u, v, \eta, \pm\infty) = \liminf_{(s,t,\zeta,\xi) \to (u,v,\eta,\pm\infty)} \varphi(s, t, \zeta, \xi) = \begin{cases} 1 & \text{if } u = v, \\ 2 & \text{if } u \ne v,\ uv \ge 0, \\ 3 & \text{if } u \ne v,\ uv < 0. \end{cases}$$

Condition (i) of Theorem 3.1 is verified, while condition (ii) is not. In fact, if $u \ne v$ and $uv < 0$, we have

$$\varphi(u, v, \xi, \eta) = 3 > 2 = \varphi(u, u, \xi, +\infty) + \varphi(v, v, +\infty, \eta).$$

Nevertheless, we can show that F, given by (3.1), is lower semicontinuous on $\mathcal{H}^2(I)$. In fact, let us take a sequence (u_h) converging to u weakly in $\mathcal{H}^2(I)$. We have to show that for every $x \in (S_u \cup S_{u'})$ we have

$$\begin{aligned}&\varphi(u(x_-), u(x_+), u'(x_-), u'(x_+))\\ &\qquad \le \liminf_{h\to+\infty} \sum_{i=1}^{m} \varphi(u_h(x^i_{h-}), u_h(x^i_{h+}), u'_h(x^i_{h-}), u'_h(x^i_{h+})),\end{aligned} \tag{4.1}$$

where (x^i_h) are the points of $S_{u_h} \cup S_{u'_h}$ that converge to x (we can suppose m independent of h). If $u(x_-)u(x_+) \ge 0$ then $\varphi(u(x_-), u(x_+), u'(x_-), u'(x_+)) \le 2$, and it is easy to see, by the lower semicontinuity of φ and condition (i) of Theorem 3.1, that (4.1) is satisfied. Then the case $u(x_-)u(x_+) < 0$ remains, so that

$$\varphi(u(x_-), u(x_+), u'(x_-), u'(x_+)) = 3.$$

If $m \ge 3$ then inequality (4.1) is trivially verified, so it is also verified by lower semicontinuity if $m = 1$. Then we can confine our analysis to the case $m = 2$ and suppose $x^1_h < x^2_h$ for every $h \in \mathbf{N}$. If we have $u(x^i_{h-}) \ne u(x^i_{h+})$ for $i = 1$ or $i = 2$ and for all $h \in \mathbf{N}$, then the inequality is satisfied again. Then the only case left out is $u_h(x^i_{h-}) = u_h(x^i_{h+})$ for $i = 1, 2$. By Remark 2.4, we must have

$$u'_h(x^1_{h+}) - u'_h(x^2_{h-}) \to 0$$

as $h \to +\infty$ and, in particular,

$$u'_h(x^1_{h+}) - u'_h(x^2_{h-}) \notin 1 + 2\mathbf{Z}.$$

Since we definitely have $u_h(x^1_h)u_h(x^2_h) < 0$, we get

$$\begin{aligned}\varphi(u(x_-), u(x_+), u'(x_-), u'(x_+)) = 3 \ &\le \varphi(u_h(x^1_{h-}), u_h(x^1_{h+}), u'_h(x^1_{h-}), u'_h(x^1_{h+}))\\ &\quad + \varphi(u_h(x^2_{h-}), u_h(x^2_{h+}), u'_h(x^2_{h-}), u'_h(x^2_{h+})),\end{aligned}$$

and hence the thesis. □

We conclude the section with another example that shows that in the general case of φ lower semicontinuous, but depending effectively on all its arguments, not even the subadditivity condition (i) of Theorem 3.1 is necessary.

Example 4.4. Let us consider the function $\varphi : \mathbf{R}^4 \to \{1, 3\}$, defined by setting

$$\varphi(u, v, \xi, \eta) = \begin{cases} 1 & \text{if } (u, \xi) = (0, 1) \text{ or } (v, \eta) = (0, 1),\\ 3 & \text{otherwise.}\end{cases}$$

The function φ is lower semicontinuous but it does not satisfy condition (i) of Theorem 3.1. In fact we have, for example,

$$\varphi(2, 1, 1, 1) = 3 > 2 = \varphi(2, 0, 1, 1) + \varphi(0, 1, 1, 1).$$

Let us consider the functional F given by (3.1). In order to prove that it is lower semicontinuous with respect to the weak convergence of $\mathcal{H}^2(I)$, it is sufficient to take $u \in \mathcal{H}^2(I)$ and consider the case of a single point $x \in S_u \cup S_{u'}$. Given a sequence (u_h) converging to u, we can suppose, as in the previous example, that there exist

exactly m sequences $(x_h^1),\dots,(x_h^m)$ of points of $S_{u_h}\cup S_{u_h'}$ converging to x. We must check (4.1). If $m\ge 3$ this inequality is trivial, so it is also trivial if $(u(x_+),u'(x_+))=(0,1)$ or $(u(x_-),u'(x_-))=(0,1)$; if $m=1$ then it is an immediate consequence of the lower semicontinuity of φ. This leaves the case $m=2$ and $(u(x_+),u'(x_+))\neq(0,1)\neq(u(x_-),u'(x_-))$. We can suppose $x_h^1<x_h^2$ for every $h\in\mathbf{N}$. Since we have $(u_h(x_{h-}^1),u_h'(x_{h-}^1))\to(u(x_-),u'(x_-))$ and $(u_h(x_{h+}^2),u_h'(x_{h+}^2))\to(u(x_+),u'(x_+))$, inequality (4.1) is violated only if

$$u_h(x_{h+}^1)=u_h(x_{h-}^2)=0,\qquad u_h'(x_{h+}^1)=u_h'(x_{h-}^2)=1.$$

But then, by Lemma 2.3, we would have

$$\int_{x_h^1}^{x_h^2}|u_h''|^2\,dt\ge\frac{6}{x_h^2-x_h^1}.$$

Since this quantity diverges as $h\to+\infty$, contradicting the convergence of (u_h) in $\mathcal{H}^2(I)$, this case is also ruled out, and the lower semicontinuity of our functional is proven. □

5. Appendix: Some remarks on functionals on special functions of bounded variation (SBV). As noted in the introduction, functionals of the form (1.1) can be included in the framework of the space $\mathrm{SBV}(\Omega)$ of special functions of bounded variation introduced by De Giorgi and Ambrosio. In this section, we briefly recall the definition of this space, referring to [1] and [10] for more details. We also confront the formulation of minimum problems in $\mathcal{H}^2(a,b)$ and $\mathrm{SBV}(a,b)$ and underline some difficulties in the extension to problems in higher dimension.

Let Ω be a bounded open subset of $\mathbf{R}^n$ and $v\in\mathrm{BV}(\Omega)$ (a function of bounded variation in Ω; see, e.g., [11] and [13]). We define S_v as the complement of the Lebesgue points of v. We recall that we can define a measure theoretical "normal" $\nu_v(x)$ at H^{n-1} a.e. $x\in S_v$ (H^{n-1} denotes the $(n-1)$-dimensional Hausdorff measure), and v^+, v^-, the approximate values of v on both sides of S_v (defined in such a way that ν_v "points towards" v^+). The triplet (ν_v,v^-,v^+) is determined up to a change of sign of ν_v and an interchange of v^+ and v^-. We denote by ∇v the approximate gradient of v, which exists a.e. on Ω. We say that $v\in\mathrm{SBV}(\Omega)$ if $v\in\mathrm{BV}(\Omega)$ and its measure derivative Dv satisfies

$$Dv(E)=\int_E\nabla v\,dx+\int_{E\cap S_v}(v^+-v^-)\nu_v\,dH^{n-1}\tag{5.1}$$

for all Borel subset E of Ω. Note that if $v\in\mathrm{BV}(\Omega)$ then $\nabla v\in(\mathrm{L}^1(\Omega))^n$.

Remark 5.1 (weak formulation in dimension one). In the one-dimensional case $\Omega=(a,b)$ we can consider the space

$$\mathrm{SBV}^2(a,b)=\big\{u\in\mathrm{SBV}(a,b):\ u'\in\mathrm{SBV}(a,b)\big\}\tag{5.2}$$

(in dimension one we write u' in place of ∇u). Then, as in §2, we can consider functionals of the form

$$\mathcal{F}(u)=\int_{(a,b)}|u''|^2dt+\int_{(a,b)}|u-g|^2dt+\sum_{t\in S_u\cup S_{u'}}\varphi(t,u^-(t),u^+(t),u'^-(t),u'^+(t)),\tag{5.3}$$

where u'' now stands for the approximate differential of u' and we choose ν_u and $\nu_{u'}$ always equal to 1.

If $\varphi \geq c > 0$ then the space $\mathcal{H}^2(a,b)$ is exactly the space of functions in $\mathrm{SBV}^2(a,b)$ on which $\mathcal{F}$ is finite; hence the two approaches are equivalent (as already noted in [8]). The introduction of the space SBV^2 allows the weakening of the coercivity conditions on φ as in [1] by requiring only that

$$\varphi(t,u,v,\xi,\eta) \geq c_0((|v-u|^{\alpha_1} \wedge c_1) + (|\eta-\xi|^{\alpha_2} \wedge c_2)) \tag{5.4}$$

for some $0 \leq \alpha_1, \alpha_2 < 1$, and strictly positive constants c_0, c_1, c_2. In this case the functional $\mathcal{F}$ may also be finite outside the space $\mathcal{H}^2(a,b)$ and, in particular, we may have a minimizing u with infinitely many crease or jump discontinuities. This phenomenon does not seem to be natural in the framework of segmentation theory. We remark that all our results continue to hold in $\mathrm{SBV}^2(a,b)$, taking into account Lemma 4.1 in [1] in the proofs, provided we define the weak convergence $u_h \rightharpoonup u$ in $\mathrm{SBV}^2(a,b)$ as $u_h \to u$ strongly in $\mathrm{L}^1(a,b)$, $u_h' \to u'$ a.e. in (a,b), and $u_h'' \rightharpoonup u''$ weakly in $\mathrm{L}^2(a,b)$ with $\sup_h \mathcal{F}(u_h) < +\infty$.

Remark 5.2. (weak formulation in higher dimension). We can generalize the space SBV^2 to Ω bounded open subset of $\mathbf{R}^n$ by defining

$$\mathrm{SBV}^2(\Omega) = \left\{u \in \mathrm{SBV}(\Omega) : \ \nabla u \in \left(\mathrm{SBV}(\Omega)\right)^n\right\}. \tag{5.5}$$

We will show that this space is not suitable for dealing with functionals which take into account discontinuities of the first derivatives.

Let us consider the simple case of the functional in (1.2). The corresponding generalization to $\mathrm{SBV}^2(\Omega)$ is

$$\mathcal{F}(u) = \int_\Omega |\nabla^2 u|^2 \, dx + \int_\Omega |u-g|^2 \, dx + \beta H^{n-1}(S_u \cap \Omega) + \alpha H^{n-1}((S_{\nabla u} \setminus S_u) \cap \Omega), \tag{5.6}$$

($\nabla^2 u$ denotes the approximate gradient of ∇u). We will show that the sets $\{u \in \mathrm{SBV}^2(\Omega) : \ \mathcal{F}(u) \leq c\}$ are not compact for the a.e. convergence, and hence no application of the direct methods of the calculus of variations as in Proposition 2.2 is possible.

The following example is from De Giorgi: Let us consider the sequence (u_h) of $\mathrm{SBV}^2_{\mathrm{loc}}(\mathbf{R}^2)$ functions defined by

$$u_h(x) = u_h(x_1,x_2) = \begin{cases} 0 & \text{if } x \notin \bigcup_{k=1}^h B((1/k,0), 2^{-(k+1)}), \\ \frac{4^k}{k}(x_1 - \frac{1}{k}) & \text{if } x \in B((1/k,0), 2^{-(k+1)}),\ k \in \{1,\dots,h\} \end{cases}$$

($B(x,\rho)$ denotes the ball of center x and radius ρ). We have, by simple calculations,

$$\begin{aligned} \mathcal{F}(u_h) &= \int_\Omega |u_h - g|^2 \, dx + \beta H^1(S_{u_h} \cap \Omega) \\ &\leq 2\int_\Omega |u_h|^2 \, dx + 2\int_\Omega |g|^2 \, dx + \beta \sum_{k=1}^h \pi 2^{-h} \\ &\leq \frac{\pi}{8} \sum_{k=1}^h \frac{1}{k^2} + 2\int_\Omega |g|^2 \, dx + \beta\pi \leq c < +\infty. \end{aligned}$$

Moreover, the sequence u_h converges in $L^2_{loc}(\mathbf{R}^2)$ to the function u given by

$$u(x) = u_h(x_1, x_2) = \begin{cases} 0 & \text{if } x \notin \bigcup_{k=1}^{\infty} B((1/k, 0), 2^{-(k+1)}), \\ \frac{4^k}{k}(x_1 - \frac{1}{k}) & \text{if } x \in B((1/k, 0), 2^{-(k+1)}),\ k = 1, 2, \ldots. \end{cases}$$

We see that by taking, for example, $\Omega = B(0, 2)$, we have

$$\int_\Omega |\nabla u| dx = \sum_{h=1}^{\infty} \frac{4^h}{h} \pi (2^{-(h+1)})^2 = \frac{\pi}{4} \sum_{h=1}^{\infty} \frac{1}{h} = +\infty.$$

Hence ∇u is not integrable and we have $u \notin \mathrm{SBV}(\Omega)$ and $\nabla u \notin \big(\mathrm{SBV}(\Omega)\big)^2$; in particular, $u \notin \mathrm{SBV}^2(\Omega)$.

REFERENCES

[1] L. Ambrosio, *A compactness theorem for a special class of functions of bounded variation*, Boll. Un. Mat. Ital., B (3), 1989, pp. 857–881.

[2] L. Ambrosio and A. Braides, *Functionals defined on partitions of sets of finite perimeter, I: Integral representation and Γ-convergence*, J. Math. Pures. Appl., 69 (1990), pp. 285–305.

[3] ———, *Functionals defined on partitions of sets of finite perimeter, II: Semicontinuity, relaxation and homogenization*, J. Math. Pures. Appl., 69 (1990), pp. 307–333.

[4] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, Cambridge, Massachussets, 1987.

[5] G. Bouchitté and G. Buttazzo, *New lower semicontinuity results for nonconvex functionals defined on measures*, Nonlinear Anal., 15 (1990), pp. 679–692.

[6] A. Braides and A. Coscia, *A singular perturbation approach to problems in fracture mechanics*, Math. Models Methods Appl. Sci., 3 (1993), pp. 303–340.

[7] G. Buttazzo, *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variations*, in Pitman Res. Notes Math. Ser. 207, Longman, Harlow, 1989.

[8] A. Coscia, *Existence results for a new variational problem in one-dimensional segmentation theory*, Ann. Univ. Ferrara Sez. VII (N.S.), 37 (1991), pp. 185–203.

[9] G. Dal Maso, *An Introduction to Γ-convergence*, Birkhäuser, Boston, 1993.

[10] E. De Giorgi and L. Ambrosio, *Un nuovo tipo di funzionale del calcolo delle variazioni*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 82 (1988), pp. 199–210.

[11] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Basel, 1983.

[12] D. Mumford and J. Shah, *Optimal approximation by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.

[13] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, 1989.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1199–1224, September 1995

007

GLOBAL ASYMPTOTIC DYNAMICS OF GRADIENT-LIKE BISTABLE EQUATIONS*

KONSTANTIN MISCHAIKOW†

Abstract. The dynamics on the global attractors of bistable gradient-like evolution equations are described via a semiconjugacy onto the dynamics of a simple system of ordinary differential equations. The fact that the semiconjugacy is "onto" implies that, given any solution to the ordinary differential equation, there exists a corresponding orbit on the attractor of the evolution equation. It is also shown that these results apply to the damped wave equation, a viscoelastic beam equation, the Fitz–Hugh–Nagumo equations, the Cahn–Hilliard equation, and the phase-field equations. The proofs are based on the Conley index theory.

Key words. Conley index, Cahn–Hilliard equation, Fitz–Hugh–Nagumo equation, viscoelastic beam, damped wave equation

AMS subject classifications. 35B40, 35K22

1. Introduction.

This paper characterizes the dynamics on the global attractor for *bistable* gradient-like differential equations. Although we shall purposely remain vague about the exact meaning of bistable, the archetype is the Chafee–Infante problem

$$\begin{aligned} u_t &= u_{xx} + \lambda^2(u - u^3), \qquad x \in (0,\pi), \\ u(0,t) &= u(\pi,t) = 0, \end{aligned} \tag{1}$$

where $\lambda \in [0,\infty)$. This equation and its generalizations have been extensively studied (see [5], [15], [11], [4]). For our purposes, however, the following result is of greatest interest. Let $\mathcal{A}^\lambda$ denote the global attractor for (1) and $\varphi^\lambda : \mathbf{R} \times \mathcal{A}^\lambda \to \mathcal{A}^\lambda$ denote the induced flow on the attractor; then the dynamics of φ^λ are completely understood up to topological conjugacy. To be more precise, let

$$D^P := \{z = (z_0, \dots, z_{P-1}) \mid \|z\| \le 1\} \subset \mathbf{R}^P$$

be the closed unit ball in $\mathbf{R}^P$ and $S^{P-1} = \partial D^P$ be the unit sphere. Let $\psi^P : \mathbf{R} \times D^P \to D^P$ denote the flow generated by the following system of ordinary differential equations:

$$\dot{\zeta} = Q\zeta - \langle Q\zeta, \zeta\rangle\zeta, \qquad \zeta \in S^{P-1}, \tag{2}$$

$$\dot{r} = r(1-r), \qquad r \in [0,1], \tag{3}$$

where

$$Q = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & \frac{1}{2} & & \\ \vdots & & \ddots & \\ 0 & & & \frac{1}{P} \end{bmatrix}.$$

The dynamics of ψ^P are easily understood if one realizes that (2) is obtained by projecting the linear system $\dot{z} = Qz$ onto the unit sphere. Let $\mathbf{e}_p^{\pm} = (0, \dots, \pm 1, \dots, 0)$

* Received by the editors June 23, 1993; accepted for publication January 3, 1994. This research was supported in part by National Science Foundation grant DMS-9101412.

† Center for Dynamical Systems and Nonlinear Studies, School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

be the unit vectors in the pth direction and $[\lambda]$ denote the greatest integer less than or equal to λ. Then one has the following result.

THEOREM 1.1. *There exists a homeomorphism $f : \mathcal{A}^\lambda \to D^{[\lambda]}$ such that the following diagram commutes:*

$$\begin{array}{ccc} \mathbf{R} \times \mathcal{A}^\lambda & \xrightarrow{id \times f} & \mathbf{R} \times D^{[\lambda]} \\ \downarrow{\scriptstyle \tilde{\varphi}^\lambda} & & \downarrow{\scriptstyle \psi^{[\lambda]}} \\ \mathcal{A}^\lambda & \xrightarrow{f} & D^{[\lambda]} \end{array}$$

where $\tilde{\varphi}^\lambda$ is a flow obtained from φ^λ by an order-preserving reparameterization of time. In other words, the dynamics on $\mathcal{A}^\lambda$ are conjugate to those on $D^{[\lambda]}$.

The proof of this theorem relies heavily on the following four facts:

1. The semiflow Φ^λ generated by (1) has a global compact attractor $\mathcal{A}^\lambda$. Furthermore, when restricted to this attractor, Φ^λ becomes a flow φ^λ.
2. The functional
$$V(u) = \int_0^1 \frac{u_x^2}{2} - \frac{u^2}{2} + \frac{u^4}{4} dx$$
is a *Lyapunov function* for φ^λ.
3. The set of equilibrium points $\mathcal{E}^\lambda$ are known exactly for each parameter value.
4. The flow on $\mathcal{A}^\lambda$ is Morse–Smale [1], [16].

By now it is well known that there are many systems of partial differential equations for which conditions (1) and (2) are satisfied. The third condition is problematic. The ability to compute the set equilibrium points depends highly on the type of evolution equation and the nonlinearity. Thus, if one wishes to prove a general result, one either assumes that this information is known or proposes a weaker notion of equilibrium set. We shall do the former in this paper, leaving the latter for another paper. The final condition is almost surely false in general, and one may view one of the contributions of this paper as providing a method for obtaining global dynamic information without explicitly requiring transversality.

Since our techniques are intended to be widely applicable, i.e., to any bistable gradient-like system, it is necessary to formulate the following assumptions abstractly.

(A1) $\mathcal{A}$ is a global compact attractor for a semiflow Φ on a Banach space. Furthermore, if φ denotes the restriction of Φ to $\mathcal{A}$ then φ defines a flow on $\mathcal{A}$.

(A2) Under the flow $\varphi : \mathbf{R} \times \mathcal{A} \to \mathcal{A}$
$$\mathcal{M}(\mathcal{A}) = \{M(p^\pm) \mid p = 0, \dots, P-1\} \cup \{M(P)\}$$
with ordering $P > P-1^\pm > \cdots > 1^\pm > 0^\pm$ being a Morse decomposition of $\mathcal{A}$.

(A3) The cohomology Conley indices of the Morse sets are
$$CH^k(M(P)) \approx \begin{cases} \mathbf{Z} & \text{if } k = P, \\ 0 & \text{otherwise.} \end{cases}$$

and for $p = 0, \dots, P-1$,

$$CH^k(M(p^{\pm})) \approx \begin{cases} \mathbf{Z} & \text{if } k = p, \\ 0 & \text{otherwise.} \end{cases}$$

(A4) The connection matrix for $\mathcal{M}(\mathcal{A})$ is given by

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ D_0 & 0 & 0 & & \vdots \\ & D_1 & & \ddots & 0 \\ \vdots & & & 0 & 0 \\ 0 & & \dots & D_{P-1} & 0 \end{bmatrix},$$

where, up to a choice of orientation for $p = 0, \dots, P-2$,

$$D_p : CH^p(M(p^-)) \oplus CH^p(M(p^+)) \to CH^{p+1}(M(p{+}1^-)) \oplus CH^{p+1}(M(p{+}1^+))$$

is given by

$$D_p = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$$

and

$$D_{P-1} : CH^{P-1}(M(P-1^-)) \oplus CH^{P-1}(M(P-1^+)) \to CH^P(M(P))$$

is given by

$$D_P = [1, -1].$$

The terms used in these assumptions are defined in §4. For the moment, however, it is sufficient to accept that these assumptions parallel the crucial facts used to understand the Chafee–Infante problem. In particular, it is obvious that (A1) replaces condition (1). (A2) and (A3) together are generalizations of (2) and (3), and (A4) will be used in place of the Morse–Smale condition. With these assumptions we obtain the following theorem.

THEOREM 1.2. *Given assumptions* (A1)–(A4) *there exist a continuous surjective map* $f : \mathcal{A} \to D^P$ *and a flow* $\tilde{\varphi}$ *obtained by an order-preserving time reparameterization of* φ *such that the following diagram commutes:*

$$\begin{array}{ccc} \mathbf{R} \times \mathcal{A} & \xrightarrow{id \times f} & \mathbf{R} \times D^P \\ \downarrow \tilde{\varphi} & & \downarrow \psi^P \\ \mathcal{A} & \xrightarrow{f} & D^P \end{array}$$

where $M(p^{\pm}) = f^{-1}(\mathbf{e}_p^{\pm})$ *for* $0 \le p \le P-1$, *and* $M(P) = f^{-1}(0)$.

Observe that the difference between Theorems 1.1 and 1.2 is that the latter only claims the existence of a semiconjugacy with the flow ψ^P. The proof of Theorem 1.2

is presented in §§6 and 7. It is based on similar work by McCord and Mischaikow [20], which described the global dynamics for scalar delay equations with negative feedback.

It is, perhaps, appropriate to comment at this point upon the reparameterized flow $\tilde{\varphi}$. It is obtained via an order-preserving time reparameterization, and hence shares the same qualitative properties as φ. We choose to make it explicit for two reasons: it is used in our proof and it allows one to decouple the space and time parameters for the map from $\mathbf{R} \times \mathcal{A}$ to $\mathbf{R} \times D^P$.

As was previously indicated, the assumptions of Theorem 1.2 are stated in a rather abstract form. To be applicable it needs to be shown that (A1)–(A4) are verifiable. Therefore, in §2 we present three hypotheses, labelled (H1)–(H3), which are meant to be "computable" in practice and yet imply (A1)–(A4).

In §3 we present a variety of "bistable" gradient-like evolution equations. In particular, we discuss the one-dimensional damped wave equation, a viscoelastic beam equation, the Fitz–Hugh–Nagumo equations for special parameter values, the Cahn–Hilliard equation, and the phase field equations, and show that hypotheses (H1)–(H3) are satisfied in each case.

This paper was organized to emphasize the applications, rather than the abstract results. Therefore, we have relegated the most abstract aspects of the proofs to the end of the paper. It is hoped that this will permit the reader who is not acquainted with the Conley index theory to become familiar with the goals of the results before plunging into the details. The reader whose primary interest is in the index theory techniques may wish to read the sections in the order §1, §6, §2, §5, and finish with the applications in §3.

2. Verifying the assumptions. Given the intended applications of this paper, i.e., bistable gradient-like equations, assumptions (A1)–(A4) are more general than necessary. We hope to justify this excess generality in future papers; however, for the moment we shall concern ourselves with the problem of verifying the assumptions. In the next section we present reasonably different examples of bistable equations and show that they all satisfy the following three hypotheses.

(H1) There exists a continuous parameterized family of semiflows on a Banach space X,

$$\Phi : \mathbf{R}^+ \times X \times \Lambda \to X \times \Lambda,$$

given by

$$\Phi(t, u, \lambda) = (\Phi^\lambda(t, u), \lambda),$$

where the parameter space Λ is a compact interval in $[0, \infty)$. Furthermore, Φ has a global compact attractor $\bar{\mathcal{A}}$ and $\Phi\,|_{\bar{\mathcal{A}}}$ generates a flow

$$\varphi : \mathbf{R} \times \bar{\mathcal{A}} \to \bar{\mathcal{A}},$$

where

$$\varphi(t, u, \lambda) = (\varphi^\lambda(t, u), \lambda)$$

and φ^λ is a flow defined on

$$\mathcal{A}^\lambda := \bar{\mathcal{A}} \cap (X \times \{\lambda\}).$$

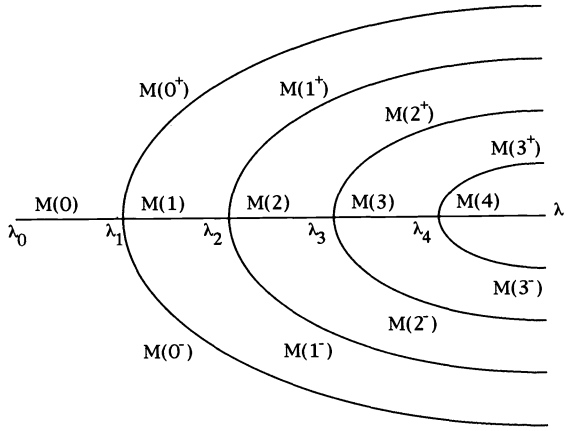


FIG. 1. *Bistable Dirichlet bifurcation diagram.*

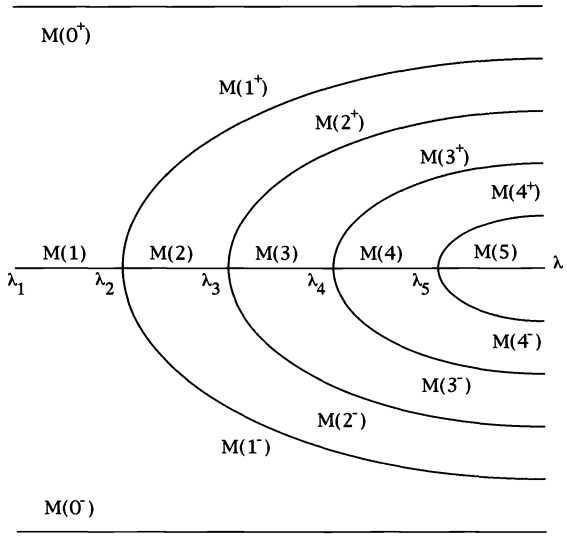


FIG. 2. *Bistable Neumann bifurcation diagram.*

(H2) As a function of λ, the equilibrium solutions $\mathcal{E}^\lambda$ of φ^λ are given by the bifurcation diagrams of Figs. 1 or 2. Furthermore, at each bifurcation point λ_p the zero solution $\mathbf{0}$ undergoes a generic supercritical pitchfork bifurcation. In the case of Fig. 2 the equilibria $M(0^\pm)$ are stable.

(H3) There exists a continuous Lyapunov function $V : \bar{\mathcal{A}} \to \mathbf{R}$ such that
  (i) for all $\lambda \in \Lambda$, if $u \notin \mathcal{E}^\lambda$ then $V(\varphi^\lambda(t,u)) < V(u)$ for all $t > 0$,
  (ii) for every p there exists v_p^λ such that $M^\lambda(p^\pm) \subset V^{-1}(v_p^\lambda)$.

The goal of this section is to show that under these hypotheses it is possible to obtain the conclusion of Theorem 1.2. In particular, we wish to prove the following theorem.

THEOREM 2.1. *Assume hypotheses* (H1)–(H3) *hold and let* $\lambda_P < \lambda < \lambda_{P+1}$. *Then, for* $\mathcal{A} = \mathcal{A}^\lambda$ *assumptions* (A1)–(A4) *are satisfied.*

While it is presumed that the meaning of hypotheses (H1)–(H3) will be clarified via the specific examples of §3, several general comments are in order at this point.

1. Most "natural" choices of parameter values give rise to continuous families of semiflows. Also, there is a rapidly growing body of literature dedicated to proving the existence of global attractors [12], [11], [24]. Therefore, (H1) can be verified for a wide range of differential equations.
2. (H2) is extremely restrictive. It is highly dependent on the choice of nonlinearity and, even if valid, can be extremely difficult to verify rigorously. On the other hand, if one is willing to accept numerical evidence, then the equilibrium sets may be computable. Furthermore, implicit in the statement that **0** undergoes a generic supercritical pitchfork bifurcation at λ_p is the assumption that, for λ close but not equal to λ_p, **0** is a hyperbolic equilibrium. In fact, in all our examples it can be shown that **0** is hyperbolic for all parameter values except the bifurcation points.
3. Lyapunov functions occur naturally in many physical models, thus (H3) should not be considered a very restrictive hypothesis.

The only nontrivial part of the proof of Theorem 2.1 is demonstrating that (A4) is satisfied. This will be done in §5. The other aspects of the proof are fairly straightforward. In fact, restricted to a specific parameter value (H1) is identical to (A1). As the following two lemmas show, (A2) and (A3) are verified almost as easily.

LEMMA 2.2. *Given* (H1)–(H3), *if* $\lambda \in (\lambda_P, \lambda_{P+1}]$, *then*

$$\mathcal{M}(\mathcal{A}^\lambda) = \{M^\lambda(p^\pm) \mid 0 \le p < P\} \cup \{M(P)\}$$

is a Morse decomposition of $\mathcal{A}^\lambda$.

Proof. By (H2), the finite set $\mathcal{M}(\mathcal{A}^\lambda)$ contains *all* the critical points in $\mathcal{A}^\lambda$. By (H3) the flow is gradient-like and, therefore, $\mathcal{M}(\mathcal{A}^\lambda)$ is a Morse decomposition (see [23, Chap. 23, §D.2]). □

Observe that (A2) is stronger than the result of this lemma. In particular, we still need to prove that $P > P - 1^\pm > \cdots > 0^\pm$ is an admissible partial ordering. Although our proof is elementary it uses some technical results from the Conley index theory, and hence, we shall present it in §5.

LEMMA 2.3. *Given* (H1)–(H3), (A3) *holds.*

Proof. We shall only present the proof in the case where the set of equilibrium points is given by Fig. 1. The argument in the setting of Fig. 2 is essentially the same and is effectively given in [23].

The proof is by induction. For $\lambda_0 \le \lambda < \lambda_1$, $\mathcal{A}^\lambda = M(0)$; therefore, by Proposition 4.1 the cohomology Conley index of $M(0)$ is

$$CH^n(M^\lambda(0)) \approx \begin{cases} \mathbf{Z} & \text{if } n = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Now, we use the fact that the bifurcation at λ_1 is a generic supercritical pitchfork bifurcation to conclude that, for λ close to but greater than λ_1, $M(1)$ is a hyperbolic fixed point with exactly one eigenvalue in the right half-plane. Hence

$$CH^n(M^\lambda(1)) \approx \begin{cases} \mathbf{Z} & \text{if } n = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, for λ close to but greater than λ_1, the two equilibrium points $M^\lambda(0^\pm)$ bifurcating out of the zero solution are asymptotically stable equilibria, and hence

$$CH^n(M^\lambda(0^\pm)) \approx \begin{cases} \mathbf{Z} & \text{if } n = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Since the branches of equilibria $M(0^{\pm})$ undergo no further bifurcations, their index remains constant. A simple induction argument now finishes the proof. □

3. Applications. Before presenting the applications, we must make a general comment. In all the examples, the spacial domain is taken as a bounded interval in $\mathbf{R}$. It is expected that these results can be extended to the setting of thin domains [13]. Obviously, for general multidimensional domains the bifurcation diagrams will be considerably more complicated than Figs. 1 and 2.

Also, note that the continuity demanded by (H1) is satisfied for all these examples, since the parameter is either the length of the spacial domain or a constant positive nonsingular coefficient.

3.1. Damped wave equation. Consider

$$u_{tt} + \alpha u_t - u_{xx} = \lambda^2 f(u), \quad x \in (0, \pi), \tag{4}$$

with Dirichlet boundary conditions

$$u(0) = u(\pi) = 0$$

and $\alpha > 0$. Assume $f \in C^2(\mathbf{R})$, $f(0) = 0$, $f'(0) = 1$, $\limsup_{|u|\to\infty} \frac{f(u)}{u} \leq 0$, and $uf''(u) < 0$ for all $u \in \mathbf{R} \setminus \{0\}$.

Then, following the discussion in [11], (H1)–(H3) are satisfied, where, as a function of λ, the bifurcation diagram is given by Fig. 1. In fact, the discussion makes it clear that the same results hold, if, in (4), the term αu_t is replaced by $h(u_t)$, where $h \in C^1(\mathbf{R}, \mathbf{R})$, $h(0) = 0$, and $0 < a \leq h' < b$ for some positive constants a and b.

3.2. One-dimensional beam with soft loading. Hattori and Mischaikow [14] considered the following model for a one-dimensional beam with soft loading:

$$u_{tt} = \sigma(u_x)_x + \nu u_{xxt} - \eta u_{xxxx}, \quad x \in (0, 1), \tag{5}$$

with boundary conditions

$$u(0, t) = 0, \quad \sigma(u_x(1, t)) + \nu u_{xt}(1, t) - \eta u_{xxx}(1, t) = P,$$

$$u_{xx}(0, t) = u_{xx}(1, t) = 0.$$

The nonlinearity $\sigma \in C^3$ given in Fig. 3 was assumed to satisfy the following growth conditions:

$$\lim_{q\to\infty} \sigma(q) \approx c_1 q^a, \qquad a \geq 1, \tag{6}$$

$$\lim_{q\to-\infty} \sigma(q) \approx c_2 q^b, \qquad b \geq 1, \tag{7}$$

at infinity, while, locally, $\sigma'(\delta) < 0$, $\sigma'''(\delta) > 0$, and $(\sigma(u + \delta) - P)/u < \sigma'(\delta)$ for $\alpha - \delta \leq u \leq \beta - \delta$ (except at $u = 0$).

A natural parameter is the capillarity coefficient η. If we let $\lambda = \frac{1}{\eta}$, then [14] or, in particular Theorem B of [14] implies that (H1)–(H3) are satisfied where the bifurcation points occur at $\lambda = (N\pi)^2/-\sigma'(\delta)$.

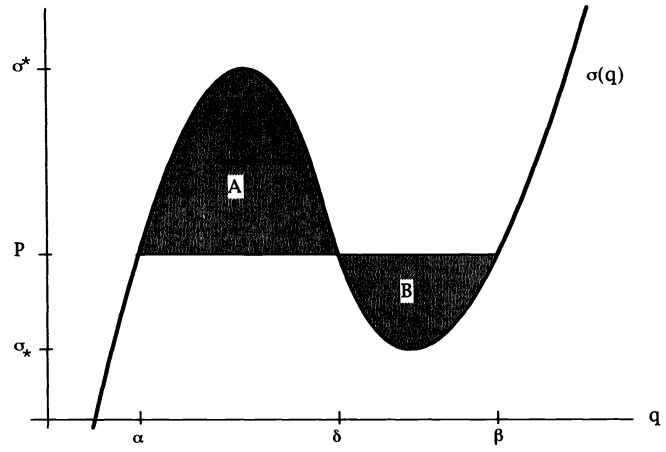


FIG. 3. *The nonlinearity* σ.

3.3. Fitz–Hugh–Nagumo equations. Conley and Smoller [7] have considered the following form of the Fitz–Hugh–Nagumo equations:

$$v_t = v_{xx} - (v-c)(v-1)v - u,$$
$$u_t = \delta v - \gamma u, \qquad (x,t) \in (-L,L) \times \mathbf{R}^+,$$

where δ and γ are positive constants and $0 < c < \frac{1}{2}$. Under the assumption of Neumann boundary conditions

$$v_x(\pm L, t) = u_x(\pm L, t) = 0,$$

they proved the following results:

1. (H1) holds because of the existence of an invariant rectangle and $\lambda = L$, the length of the domain;
2. (H2) holds where the set of equilibria as a function of L takes the form of Fig. 2.
3. (H3) is satisfied if $\gamma^2 \geq \delta$.

3.4. Cahn–Hilliard. In one space dimension the Cahn–Hilliard equation takes the form

$$u_t = (-\epsilon^2 u_{xx} + f(u))_{xx},$$

where $(x,t) \in [-1,1] \times \mathbf{R}^+$ with boundary conditions

$$u_x(\pm 1, t) = u_{xxx}(\pm 1, t) = 0.$$

The nonlinearity is usually assumed to be of cubic type with three simple zeros. This equation serves as a phenomenological model for the process of phase separation of a binary alloy at a fixed temperature. Thus, the fact that

$$\int_{-1}^{1} u(x,t)dx = M,$$

where M is a constant independent of time is essential to the model. Let $\lambda = \epsilon^{-1}$ be the parameter. The reader is referred to [24, Chap. III, §4.2] for proofs that (H1) and

(H3) are satisfied. As might be expected, (H2) has proven to be the assumption most resistant to proof. Although considerable numerical work has been done and has led to general agreement regarding the nature of the bifurcation diagram as a function of λ, there are few rigorous results. The little that is known is in the case in which

$$f(u) = u^3 - u$$

and

$$\int_{-1}^{1} u(x,t)dx = 0.$$

In this setting the results of Zheng [25] and Bates and Fife [2] imply that (H2) holds with Fig. 1 as the bifurcation diagram.

3.5. Phase-field equations. In its simplest form, the phase-field equations (see [2], [3], and [8] can be written as

$$(T + lu)_t = KT_{xx}, \tag{8}$$

$$\tau u_t = \epsilon^2 u_{xx} - u^3 + u + \sigma T, \tag{9}$$

where $(x,t) \in [-1,1] \times \mathbf{R}^+$. One can either choose Dirichlet boundary conditions

$$u(\pm 1, t) = T(\pm 1, t) = 0 \tag{10}$$

or Neumann boundary conditions,

$$u_x(\pm 1, t) = T_x(\pm 1, t) = 0; \tag{11}$$

τ, l, ϵ, K,and σ are assumed to be positive constants.

With $\lambda = \epsilon^{-1}$, Bates and Zheng [3] prove that (H1) is satisfied. (It should be noted that, in order to have a compact attractor, the space on which one considers the global flow Φ differs depending on whether one assumes Dirichlet or Neumann boundary conditions (see [3, §4]).

Checking for the validity of (H2) is equivalent to solving

$$0 = T_{xx}, \tag{12}$$

$$0 = \epsilon^2 u_{xx} - f(u) + \sigma T \tag{13}$$

with appropriate boundary conditions for all $\epsilon > 0$. We leave it to the reader to check that if one assumes Dirichlet boundary conditions, then the bifurcation diagram agrees with that of Fig. 1. The only nontrivial question is whether the bifurcations about the zero solutions are in fact generic supercritical pitchfork bifurcations. This case can be checked by studying the dispersion relation about the zero solution (see Fife [8]). The case of Neumann boundary conditions is more difficult. As in the Cahn–Hilliard equation there is a conserved quantity

$$\int_{-1}^{1} (T(x) + lu(x))\, dx = \text{constant}.$$

If we assume that $\int_{-1}^{1} (T(x) + lu(x))\, dx = 0$, then the work of Zheng [25] can again be applied to show that the bifurcation diagram agrees with Fig. 2. For a general constant the bifurcation diagram and the stability of each branch remain open questions. A general discussion of the stability properties of the equilibria of this equation can be found in [2].

4. The Conley index theory. We begin with a brief review of the relevant portions of the Conley index theory. The basic references for this material are [6], [17], [18], [21]–[23].

The Conley index was introduced to study *isolated invariant sets*, i.e., invariant sets S for which there is a compact neighborhood N of S such that S is the maximal invariant set in N. The neighborhood N is referred to as an *isolating neighborhood* for S. The Conley index of an isolated invariant set S is computed via an *index pair*, i.e., a compact pair (N, L) such that

1. $\overline{N \setminus L}$ is an isolating neighborhood for S and S is contained in the interior of N;
2. L is *positively invariant* in N, i.e., if $u \in L$ and $\varphi([0,T],u) \subset N$, then $\varphi([0,T],u) \subset L$;
3. L is an *exit set* for N, i.e., if $u \in N$ and $\varphi(T,u) \notin N$, then $\varphi(t,u) \in L$ for some $0 < t < T$.

An index pair is also said to be *regular* if, in addition, the function $\varpi : N \to [0,\infty)$ defined by

$$\varpi(u) = \begin{cases} \sup\{t > 0 | \varphi([0,t],u) \subset N \setminus L\} & \text{if } u \in N \setminus L, \\ 0 & \text{if } u \in L \end{cases}$$

is continuous. Observe that this implies that, for a regular index pair, L is a neighborhood deformation retract (along flow lines) in N. Index and regular index pairs always exist and the homotopy type of the quotient space N/L is independent of the index pair chosen. This homotopy type is the Conley index of S.

If regular index pairs are ordered by inclusion then there exists an inverse system $\{H^*(N_\alpha, L_\alpha)\}$ of index pairs with the inclusion induced cohomology map $H^*(N_\alpha, L_\alpha) \to H^*(N_\beta, L_\beta)$ an isomorphism for every $\beta < \alpha$. The inverse limit of this system, denoted $CH^*(S)$, is the cohomology Conley index of S. Since each bonding map in the system is an isomorphism, $CH^*(S) \cong H^*(N_\alpha, L_\alpha)$ for every α. In other words, the cohomology of any index pair represents the cohomology Conley index.

The following proposition [14, Thm. 6.2], [20, Cor. 3.2] provides the cohomology index for an attractor satisfying assumption (A1).

PROPOSITION 4.1. *If X is a Banach space and $\mathcal{A}$ is a global compact attractor for a continuous semiflow Φ on X, then*

$$CH^*(\mathcal{A}) \cong \begin{cases} \mathbf{Z} & \text{if } n = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Another result of this nature, which arises as a direct application of the Thom isomorphism theorem, is the following proposition.

PROPOSITION 4.2. *If S is a normally hyperbolic invariant set for a flow on a manifold with orientable unstable manifold of (normal) dimension u, then $CH^{q+u}(S) \cong H^q(S)$.*

In addition to the index, Conley introduced the concept of a Morse decomposition of an isolated invariant set. To be precise, given an isolated invariant set S, a *Morse decomposition* of S is a finite collection of disjoint compact invariant subsets of S,

$$\mathcal{M}(S) = \{M(p) \mid p \in \mathcal{P}\},$$

from which one can define a Lyapunov function; i.e., there exists a continuous function $V : S \to \mathbf{R}$ such that, if $u \notin \bigcup_{p\in\mathcal{P}} M(p)$ and $t > 0$, then $V(u) > V(\varphi(t,u))$. These

individual invariant subsets $M(p)$ are called *Morse sets* and the remaining portion $S \setminus \cup M(p)$ is referred to as the set of *connecting orbits*. In particular, given two Morse sets $M(p)$ and $M(q)$, the set of connecting orbits from $M(p)$ to $M(q)$ is defined as

$$C(M(p), M(q)) := \{u \in S \mid \omega(u) \subset M(q),\ \alpha(u) \subset M(p)\}.$$

Because of the Lyapunov function, if $C(M(p), M(q)) \neq \emptyset$ then $C(M(q), M(p)) = \emptyset$. This implies that one can impose a partial order on the indexing set $\mathcal{P}$ by setting $p > q$ if $C(M(p), M(q)) \neq \emptyset$ and taking the transitive closure. This order is called the *flow-defined order* on $\mathcal{P}$.

If $\mathcal{M}(S) = \{M(p) \mid p \in \mathcal{P}\}$ is a Morse decomposition of S, then each $M(p)$ is an isolated invariant set. S contains other isolated invariant sets, some of which can be produced by the partial order on $\mathcal{P}$ as follows. A subset $I \subset \mathcal{P}$ is an *interval* in $\mathcal{P}$ if $r \in I$ whenever $p < r < q$ and $p, q \in I$. Disjoint intervals I and J are ordered $I < J$ if $i < j$ for every $i \in I, j \in J$. They are adjacent if $IJ = I \cup J$ is also an interval (i.e., if no element of $\mathcal{P}$ lies "between" I and J). If I is an interval, let

$$M(I) := \Big(\bigcup_{i \in I} M(i)\Big) \bigcup \Big(\bigcup_{i,j \in I} C(M(j), M(i))\Big).$$

The simplest nontrivial Morse decomposition is perhaps the most important one. An *attractor-repeller pair* in S consists of two sets (A, R) such that

1. A is an *attractor* in S; i.e., there is a positively invariant neighborhood U of A in S with $\omega(U) = A$;
2. R is the *dual repeller* to A in S; i.e., $R = S \setminus \{u | \omega(u) \subset (A)\}$.

Note that A and R are both isolated invariant sets, and if

$$C(R, A) = \{u \in S \mid \alpha(u) \subset R, \omega(u) \subset A\},$$

then $S = R \cup C(R, A) \cup A$. Observe that, given a Morse decomposition and two adjacent intervals I and J in the indexing set with $I < J$, $(M(I), M(J))$ is an attractor-repeller pair for $M(IJ)$.

In an attractor-repeller decomposition, the Conley indices of the total invariant set, the attractor, and the repeller are naturally related by an *index triple*. An index triple for an attractor-repeller pair (A, R) in S is a triple of compact spaces (N, M, L) such that (N, L) is an index pair for S, (N, M) is an index pair for R, and (M, L) is an index pair for A. Such triples exist for any attractor-repeller decomposition, as do *regular* index triples, i.e., triples such that L and M are both neighborhood deformation retracts in N. In this case the cohomology exact sequence of the triple

$$\xrightarrow{\delta} H^k(N, M) \to H^k(N, L) \to H^k(M, L) \xrightarrow{\delta} H^{k+1}(N, M) \to$$

induces an exact sequence

$$\xrightarrow{\delta} CH^k(R) \to CH^k(S) \to CH^k(A) \xrightarrow{\delta} CH^{k+1}(R) \to,$$

which is known as the cohomology attractor-repeller sequence. The boundary map δ is called the *connection map*, because $\delta \neq 0$ implies that connections between R and A exist.

All of these objects have generalizations to Morse decompositions. Index triples are generalized to index filtrations and the attractor-repeller sequence is generalized

to the construction of connection matrices. Recall that the *connection matrix* is a linear map defined on the graded modules made up of the sum of the cohomology indices of Morse sets in a Morse decomposition. In our case

$$\Delta : \bigoplus_{p \in \mathcal{P}} CH^*(M(p)) \to \bigoplus_{p \in \mathcal{P}} CH^*(M(p)).$$

Furthermore, connection matrices satisfy the following conditions:

1. They are *lower triangular*; i.e., if $p \not> q$ then $\Delta(q,p) = 0$.
2. They are *coboundary operators*; i.e., they are degree $+1$ maps

$$\Delta(q,p)CH^n(M(q)) \subset CH^{n+1}(M(p))$$

and they square to zero; $\Delta \circ \Delta = 0$.

3. If p and q are adjacent in the flow-defined order then the connection matrix entry $\Delta(q,p)$ equals the connecting homomorphism for the attractor-repeller pair $(M(q), M(p))$ of $M(qp)$.
4. The relation between the local cohomology indices, i.e., that of the Morse sets, and the global cohomology index is

$$CH^*(S) \approx \frac{\ker \Delta}{\text{image} \Delta}.$$

The following theorem from Franzosa [9] is fundamental.

THEOREM 4.3. *Given a Morse decomposition, there exists at least one connection matrix.*

A new feature of the cohomology index, introduced in [20], is a pairing of the cohomology Conley index of an invariant set and the Cech cohomology of the invariant set. If (N, L) is an index pair for an isoloated invariant set S, the cup product defines a pairing

$$H^p(N) \otimes H^q(N, L) \to H^{p+q}(N, L).$$

Since the collection

$$\{N_\alpha \mid (N_\alpha, L_\alpha) \text{ is an index pair}\}$$

is cofinal with the set of neighborhoods of S, this pairing defines a pairing

$$\check{H}^p(S) \otimes CH^q(S) \to CH^{p+q}(S).$$

This pairing exists for any invariant set in any flow. If $T_I \in CH^n(M(I))$ is a generator, then there is a map

$$\iota_I^n : \check{H}^p(M(I)) \to CH^{p+n}(M(I))$$

defined by

$$\iota_I^n(z) = z \cup T_I.$$

Another aspect of the index which we shall make use of is its behavior under semiconjugacies (cf. [17], [18]). The essence of the matter is that the index theory is natural with respect to semiconjugacies as long as one works with preimages rather

than images. A technicality is that the semiconjugacy must be a *proper map*; i.e., preimages of compact sets must be compact. Thus, if $f : X \to Y$ is a proper semiconjugacy and S is an isolated invariant set in Y with index pair (N, L), then $T = f^{-1}(S)$ is an isolated invariant set in X with index pair $(f^{-1}(N), f^{-1}(L))$. Therefore, there are maps $f_* : CH_*(T) \to CH_*(S)$ and $f^* : CH^*(S) \to CH^*(T)$. The pairing defined above commutes with this map; i.e., there is a commutative diagram

$$\begin{array}{ccc} \check{H}^p(S) \otimes CH^q(S) & \to & CH^{p+q}(S) \\ \downarrow f^* \otimes f^* & & \downarrow f^* \\ \check{H}^p(T) \otimes CH^q(T) & \to & CH^{p+q}(T). \end{array}$$

Similarly, if $\{M(p)\}$ is a Morse decomposition of S then $\{T(p) = f^{-1}(M(p))\}$ is a Morse decomposition of T, and any admissible ordering on S gives an admissible ordering on T. Thus we can use the same ordering for both decompositions, and if I is an interval in that ordering, there is a map $CH^*(M(I)) \to CH^*(T(I))$. Moreover, the attractor-repeller sequence is natural: if I and J are adjacent intervals with $I < J$ then there is a commutative diagram

$$\begin{array}{ccccccc} \xrightarrow{\delta} & CH^p(M(J)) & \to & CH^p(M(IJ)) & \to & CH^p(M(J)) & \to \\ & \downarrow f^* & & \downarrow f^* & & \downarrow f^* & \\ \xrightarrow{\delta} & CH^p(T(J)) & \to & CH^p(T(IJ)) & \to & CH^p(T(J)) & \to . \end{array}$$

As was indicated in the introduction, our theorem can be stated more clearly if we use a time-reparameterized flow $\tilde{\varphi}$ rather that the actual flow φ. The details of constructing the reparameterization are uninteresting but can be found in [20, §5]. Therefore, here we merely assert the existence of the following result.

PROPOSITION 4.4. *Given a flow $\varphi : \mathbf{R} \times \mathcal{A} \to \mathcal{A}$ satisfying* (A2) *there exists:*

(i) *$\tilde{\varphi} : \mathbf{R} \times \mathcal{A} \to \mathcal{A}$, a flow obtained via an order-preserving reparameterization of time;*

(ii) *sets $N_{p^\pm}, L^+_{p^\pm}, L^-_{p^\pm}$ for $p = 0, \dots, P$ (let $P = P^+$) such that*

(a) *$N_{p^\pm}$ are isolating neighborhoods of $M(p^\pm)$, respectively, and $N_{p^+} \cap N_{p^-} = \emptyset$;*

(b) *$\partial N_{p^\pm} = L^+_{p^\pm} \cup L^-_{p^\pm}$;*

(c) *$L^\pm_{p^\pm}$ are local sections of $\tilde{\varphi}$;*

(d) *$(N_{p^\pm}, L^-_{p^\pm})$ is an index pair for $M(p^\pm)$ under $\tilde{\varphi}$;*

(e) *$(N_{p^\pm}, L^+_{p^\pm})$ is an index pair for $M(p^\pm)$ under $\tilde{\varphi}'$, where $\tilde{\varphi}'(t, u) = \tilde{\varphi}(-t, u)$;*

(f)

$$\tilde{\varphi}(\mathbf{R}, u) \cap N_{p^\pm} = \tilde{\varphi}(I_{p^\pm}(u), u) \cap N_{p^\pm},$$

where $I_{p^\pm}$ is a closed interval;

(iii) *a Lyapunov function*

$$\tilde{V} : \mathcal{A} \to [0, P]$$

such that

(a) *if $u \in M(p^\pm) \cup \left(L^+_p \cap L^-_p\right)$ then $\tilde{V}(u) = p$;*

(b) *if*

$$\tilde{\varphi}([0, t], u) \bigcap \bigcup_{p=0}^{P} N_{p^\pm} = \emptyset$$

then

$$\tilde{V}(\tilde{\varphi}(t,u)) = \tilde{V}(u) - t.$$

Since this proposition guarantees that $I_{p^\pm}(u)$ is a closed interval, we write

$$I_{p^\pm}(u) = [a_{p^\pm}(u), b_{p^\pm}(u)]$$

with the understanding that if $I_{p^\pm}(u) = \emptyset$ then $a_{p^\pm}(u)$ and $b_{p^\pm}(u)$ are not defined, and if $I_{p^\pm}(u)$ is unbounded then $a_{p^\pm}(u) = -\infty$ and/or $b_{p^\pm}(u) = \infty$. Let

$$\Theta_{p^\pm} := \{x \in \mathcal{A} | \, I_{p^\pm}(u) \neq \emptyset\};$$

then the fact that $(N_{p^\pm}, L^\pm_{p^\pm})$ are regular index pairs gives rise to the following lemma.

LEMMA 4.5. *The functions* $a_{p^\pm}, b_{p^\pm} : \Theta_{p^\pm} \to [-\infty, \infty]$ *are continuous.*

5. Proof of Theorem 2.1. The purpose of this section is to conclude the proof of Theorem 2.1. Recall that in §2 it was shown that hypotheses (H1)–(H3) implied assumptions (A1), (A3), and the fact that $\mathcal{M}(\mathcal{A}) = \{M(p^\pm) \mid p = 0, 1, \ldots, P-1\} \cup \{M(P)\}$ was a Morse decomposition for $\mathcal{A}$. It remains to be shown that the partial order of (A2) is admissible and (A4) holds. Therefore, we begin with the following lemma.

LEMMA 5.1. *The partial order*

$$P > P - 1^\pm > \cdots > 1^\pm > 0^\pm$$

is an admissible order for $\mathcal{M}(\mathcal{A}^\lambda)$.

Proof. We begin with the observation that the connection matrix must be of the form

$$\Delta^\lambda = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ D_0 & 0 & & & \vdots \\ \vdots & D_1 & & \ddots & 0 \\ & & & 0 & 0 \\ 0 & & & D_{P-1} & 0 \end{bmatrix}, \tag{14}$$

where, for $p = 1, \ldots, P-1$,

$$D_p : CH^p(M^\lambda(p^-)) \oplus CH^p(M^\lambda(p^+)) \to CH^{p+1}(M^\lambda(p+1^-)) \oplus CH^{p+1}(M^\lambda(p+1^+))$$

and

$$D_{P-1} : CH^{P-1}(M^\lambda(P-1^-)) \oplus CH^{P-1}(M^\lambda(P-1^+)) \to CH^P(M^\lambda(P)).$$

By (A3),

$$D_p : \mathbf{Z} \oplus \mathbf{Z} \to \mathbf{Z} \oplus \mathbf{Z}$$

and

$$D_{P-1} : \mathbf{Z} \oplus \mathbf{Z} \to \mathbf{Z}.$$

By (A1) and Proposition 4.1, for $n \geq 1$,

$$0 \approx CH^n(\mathcal{A}^\lambda) \approx \frac{\ker D_n}{\text{image} D_{n-1}}.$$

Therefore, the rank of D_n equals the dimension of the kernel of D_n for all $p = 0, \ldots, P-1$. Now, the image of D_{P+1} equals 0, which implies that $\ker D_P = 0$ or, equivalently, the rank of D_P is 1. In particular, we obtain that

$$\text{rank } D_p = 1$$

for all $p = 1, \ldots, P$. Since the connection matrix is strictly lower triangular and $V(M(p^+)) = V(M(p^-))$, an admissible order is

$$P > P - 1^\pm > \cdots > 1^\pm > 0^\pm. \qquad \square$$

For the remainder of this section ϵ will denote a positive but sufficiently small real number.

The proof that (A4) holds is effectively a proof by induction and consists of two distinct parts:

(1) assuming that Δ^{λ_P} is known, and, thereby, computing $\Delta^{\lambda_P+\epsilon}$;

(2) showing that $\Delta^{\lambda_P+\epsilon} = \Delta^\lambda$ for all $\lambda \in (\lambda_P, \lambda_{P+1}]$.

Part (2) follows from [10] or [14, Lem. 5.12]. The rest of this section is devoted to the proof of part (1) and is based on the argument of [14].

Before we can begin the induction step we need to compute Δ^λ for $\lambda \in [\lambda_0, \lambda_3]$. We start by observing that if $\lambda \in [\lambda_0, \lambda_1]$ then $\mathcal{A} = M(0)$, and hence

$$\Delta^\lambda = [0].$$

So let us assume that $\lambda \in (\lambda_1, \lambda_2]$. By equation (14),

$$\Delta^\lambda = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \alpha & \beta & 0 \end{bmatrix}.$$

For $\lambda = \lambda_2 - \epsilon$, $M(1)$ is a hyperbolic fixed point. Thus, for this value of λ, (A3) implies that $M(1)$ has a one-dimensional unstable manifold. Since $M(0^\pm)$ are attracting fixed points, the unstable manifold of $M(1)$ intersects transversely with the stable manifold of $M(0^\pm)$. Therefore, by [19], α and β denote the number of connecting orbits from $M(1)$ to $M(0^+)$ and $M(0^-)$, respectively. Since $\mathcal{A}$ is connected [11], $(\alpha, \beta) = (\approx, \approx)$. Now, by an appropriate choice of orientation for the generator of $CH^*(M(1))$, we can assume $(\alpha, \beta) = (1, -1)$. As was mentioned above, the results of [10] imply that

$$\Delta^\lambda = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & -1 & 0 \end{bmatrix}$$

for all $\lambda \in (\lambda_1, \lambda_2]$.

We now compute $\Delta^{\lambda_2+\epsilon}$. Again, by equation (14),

$$\Delta^{\lambda_2+\epsilon} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \alpha & \gamma & 0 & 0 & 0 \\ \beta & \delta & 0 & 0 & 0 \\ 0 & 0 & \zeta & \eta & 0 \end{bmatrix}.$$

By Lemma 5.1, the set $\{2, 1^{\pm}\}$ is an interval in $\{2, 1^{\pm}, 0^{\pm}\}$, and hence $M(2, 1^{\pm})$ is an isolated invariant set with Morse decomposition

$$\{M(2), M(1^{+}), M(1^{-})\}.$$

The connection matrix for this Morse decomposition is the submatrix of $\Delta^{\lambda_2+\epsilon}$ given by

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \zeta & \eta & 0 \end{bmatrix}.$$

Since $0 < \epsilon << 1$, the entries ζ and η are determined by the pitchfork bifurcation. In particular, we can assume transverse intersections between the stable and unstable manifolds, and hence, as in the aforementioned argument, an appropriate choice of orientation of the generator of $CH^*(M(2))$ implies that $(\zeta, \eta) = (1, -1)$. By definition $\Delta^{\lambda_2+\epsilon} \circ \Delta^{\lambda_2+\epsilon} = 0$. Thus $\alpha = \beta$ and $\gamma = \delta$. Again, for ϵ small, $M^{\lambda_2+\epsilon}(1^{\pm})$ are hyperbolic fixed points, and hence, repeating the argument used to compute $\Delta^{\lambda_2-\epsilon}$, we can assume that $\alpha = 1$ and $\gamma = -1$.

We are now in a position to perform the induction step. So let us assume that (A4) holds for all $\lambda \in [\lambda_0, \lambda_P]$; we need to show that it holds for $\lambda_P + \epsilon$. By Lemma 5.1 the set $\{P, P-1^{\pm}\}$ is an interval in $\mathcal{P}$, and hence $M(P, P-1^{\pm})$ is an isolated invariant set with Morse decomposition

$$\{M(P), M(P-1^{+}), M(P-1^{-})\}.$$

The connection matrix for this Morse decomposition is the submatrix of $\Delta^{\lambda_2+\epsilon}$ given by

$$\begin{bmatrix} 0 & 0 \\ D_{P-1} & 0 \end{bmatrix}.$$

As before, the assumption of a generic supercritical pitchfork bifurcation implies that

$$D_{P-1} = [1, -1].$$

Again as before, this forces

$$D_{P-2} = \begin{bmatrix} \alpha & \gamma \\ \alpha & \gamma \end{bmatrix}.$$

Changing our focus for a moment, by Lemma 5.1 the set $\{0^{\pm}, 1^{\pm}, \ldots, P-2^{\pm}\}$ is an interval in $\mathcal{P}$, and hence $M(0^{\pm}, 1^{\pm}, \ldots, P-2^{\pm})$ is an isolated invariant set with Morse decomposition

$$\{M(0^{+}), M(0^{-}), \ldots, M(P-2^{+}), M(P-2^{-})\}.$$

The connection matrix for this Morse decomposition is the submatrix of $\Delta^{\lambda_2+\epsilon}$ given by

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ D_0 & 0 & & & \vdots \\ \vdots & D_1 & & \ddots & 0 \\ & & & 0 & 0 \\ 0 & & & D_{P-2} & 0 \end{bmatrix}. \tag{15}$$

Now, observe that the structure of this invariant set cannot be affected by the pitchfork bifurcation. The easiest way to see this is to relabel the invariant set $M(P, P-1^{\pm})$ by $M(\bar{P})$ and consider the connection matrix for the Morse decomposition

$$\mathcal{M}(\mathcal{A}) = \{M(p^{\pm}) \mid p = 0, 1, \ldots, P-1\} \cup \{M(\bar{P})\}.$$

The same argument that proves part (2) also guarantees that the connection matrix for this Morse decomposition equals Δ^{λ_P}. Therefore, by the induction hypothesis

$$D_{P-2} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}.$$

At this point we have computed D_p for all $p \neq P-1$. To finish the computation recall that $\Delta^{\lambda_2+\epsilon} \circ \Delta^{\lambda_2+\epsilon} = 0$. This implies that $D_{P-2} \circ D_{P-1} = 0$, and hence $\alpha = -\gamma$.

Therefore, (A4) is implied by (H1)–(H3), and hence Theorem 2.1 is proven.

As a final remark, this also proves that $P > P-1^{\pm} > \cdots > 0^{\pm}$ is the flow-defined order.

6. The semiconjugacy. The construction of the map $f : \mathcal{A} \to D^P$ consists of four steps.

Step 1. Define a function

$$\hat{f} : \mathcal{A} \to \partial\left([0, P] \times [-1, 1]^P\right).$$

Step 2. Define an equivalence relation $\sim$ on $\partial\left([0, P] \times [-1, 1]^P\right)$ along with the corresponding quotient map

$$\mathcal{Q} : \partial\left([0, P] \times [-1, 1]^P\right) \to \frac{\partial\left([0, P] \times [-1, 1]^P\right)}{\sim}$$

and then show that $\tilde{f} := \mathcal{Q} \circ \hat{f}$ is continuous.

Step 3. Show that $\tilde{f} : \mathcal{A} \to D^P$.

Step 4. Show that under the image of $\hat{f}$, $\widetilde{\varphi}$ induces a flow $\widetilde{\psi}$ on D^P, and there exists a conjugacy $g : D^P \to D^P$ between $\widetilde{\psi}$ and ψ. Finally, define

$$f = g \circ \tilde{f}.$$

This function f is the semiconjugacy of Theorem 1.2, though it is left to the next section to show that f is onto.

Before proceeding with the details we shall indicate why the above steps are natural. To describe the dynamics on $\mathcal{A}$ in terms of the Morse decomposition, there are several obvious parameters which should be included.

1. For each Morse set $M(p^*)$ ($* = +$ or $-$) one would like to know how "close" the orbit of $u \in \mathcal{A}$ passes by $M(p^*)$. Thus we shall define a function

$$\tau_p : \mathcal{A} \to [-1, 1], \qquad p = 1, \ldots, P-1$$

with the properties that $\tau_p(u) = 1$ implies that $\omega(u)$ or $\alpha(u) \subset M(p^+)$, $\tau_p(u) = -1$ implies that $\omega(u)$ or $\alpha(u) \subset M(p^-)$, while $\tau_p(u) = 0$ implies that the orbit of u does not intersect the interior of N_p, the isolating neighborhood of $M(p)$.

2. Given two points u and u' on the same orbit, $\tau_p(u) = \tau_p(u')$. To distinguish these points we shall make use of a Lyapunov function

$$V : \mathcal{A} \to [0, P].$$

Given these functions, we define

$$\begin{aligned}\hat{f}(u) &= (V(u), \tau(u)) \\ &= (V(u), \tau_0(u), \ldots, \tau_{P-1}(u)) .\end{aligned}$$

It is easily seen that $\hat{f}$ cannot be continuous. For example, if one considers a sequence of points whose omega limit set is in $M(p^+)$, then $\tau_p = 1$ on this sequence. However, this sequence may limit to a point whose omega limit set is in $M(q^*)$, where $q > p$, in which case $\tau_p = 0$ for the limit point. The quotient performed in Step 2 addresses this problem; i.e., the resulting map $\tilde{f}$ is continuous.

In Step 3 it will be shown that $\tilde{f} : \mathcal{A} \to D^P$. This is not immediate from the work of Steps 1 and 2 and requires a better understanding of $\hat{f}(\mathcal{A})$. However, it does provide a good opportunity for further exploration of the relationship between the model flow $\psi : \mathbf{R} \times D^P \to D^P$ and the construction of the semiconjugacy.

Finally, it needs to be shown that, under $\tilde{f}$, $\tilde{\varphi}$ induces a flow $\tilde{\psi}$ on D^P. Then the results of Step 3 can be used to argue that $\tilde{\psi}$ is conjugate to ψ via a map g, and hence $f = g \circ \tilde{f}$ is the desired semiconjugacy.

6.1. The map $\tilde{f} : \mathcal{A} \to \partial([0, P] \times [-1, 1]^P)$. As was mentioned before, the functions

$$\tau_p : \mathcal{A} \to [-1, 1]$$

are intended to provide information on how trajectories in $\mathcal{A}$ pass by the Morse sets $M(p^{\pm})$. Guided by the idea that the longer the trajectory stays in the isolating neighborhood the closer it is to the Morse set, define

$$\lambda_{p^{\pm}}(u) = \begin{cases} \infty & \text{if } b_{p^{\pm}}(u) = \infty \text{ or } a_{p^{\pm}}(u) = -\infty, \\ 0 & \text{if } I_{p^{\pm}}(u) = \emptyset, \\ b_{p^{\pm}}(u) - a_{p^{\pm}}(u) & \text{otherwise,} \end{cases} \tag{16}$$

using the functions $b_{p^{\pm}}$, $a_{p^{\pm}}$, and $I_{p^{\pm}}$ defined in §4. Set

$$\tau_p(u) = \frac{2}{\pi} \tan^{-1}(\lambda_{p^+}(u) - \lambda_{p^-}(u)), \tag{17}$$

where $\tan^{-1}(\pm\infty) = \pm\frac{\pi}{2}$.

LEMMA 6.1. *For $p = 0, \ldots, P-1$,*

$$\tau_p : \Theta_p \to [-1, 1]$$

is a continuous function. Furthermore,

(i) $\tau_p(u) = 1$ *if and only if* $\omega(u)$ *or* $\alpha(u)$ *is contained in* $M(p^+)$;

(ii) $\tau_p(u) = -1$ *if and only if* $\omega(u)$ *or* $\alpha(u)$ *is contained in* $M(p^-)$;

(iii) $\tau_p(u) = 0$ *if and only if* $\tilde{\varphi}(\mathbf{R}, u) \cap N_{p^{\pm}} = \emptyset$ *or* $\tilde{\varphi}(\mathbf{R}, u) \cap N_{p^*} \subset L_{p^{\pm}} \cap L^{\pm}_{p^{\pm}}$.

Proof. By Lemma 4.5, $a_{p^{\pm}}$ and $b_{p^{\pm}}$ are continuous functions. Thus $\lambda_{p^{\pm}}$ and τ_p are continuous.

(i) and (ii) hold since $|\tau_p(u)| = 1$ is equivalent to $|\lambda_{p^{\pm}}(u)| = \infty$.

Finally, as is obvious from (16), if $\varphi(\mathbf{R}, u) \cap N_{p^{\pm}}(u) = \emptyset$ then $\lambda_{p^{\pm}} = 0$. Furthermore, by Proposition 4.4, if $\lambda_{p^+}(u) \neq 0$, then $\lambda_{p^-}(u) = 0$ and vice versa. Thus, $\tau_p(u) = \lambda_{p^+}(u)$ or $-\lambda_{p^-}(u)$ depending on whether $\varphi(\mathbf{R}, u)$ intersects N_{p^+} or N_{p^-}. □

Turning to the definition of the Lyapunov function, we begin by defining local Lyapunov functions

$$V_p : N_p \to \left[p - \frac{1}{4}, p + \frac{1}{4}\right]$$

by

$$V_p(u) = \begin{cases} p & \text{if } u \in M_p, \\ W_p^+(\widetilde{\varphi}(a_p(u), u)) + \frac{1}{2\pi}\tan^{-1}\left(\frac{a_p(u)}{2}\right) & \text{if } a_p(u) > -\infty, \\ W_p^-(\widetilde{\varphi}(b_p(u), u)) + \frac{1}{2\pi}\tan^{-1}\left(\frac{b_p(u)}{2}\right) & \text{if } b_p(u) < \infty. \end{cases}$$

To define V from the isolating neighborhoods we make use of the Lyapunov function $\widetilde{V}$ of Proposition 4.4. Let

$$K_p := \{u \in \mathcal{A} \mid \tilde{V}(u) = p\} \setminus N_p$$

and define

$$K_p^+ := K_p \cup L_p^+.$$

Observe that K_p^+ is a section for the flow. Thus, if $x \in \mathcal{A} \setminus \bigcup_{p=0}^{P} N_p$, then there exists a unique $k_u \in K_{p_u}^-$, where $p_u = 0, 1, \ldots, P-1$, and a unique $\alpha_u \in [0, 1)$ such that

$$\widetilde{\varphi}(\alpha_u, k_u) = u.$$

Now define

$$V(u) = \begin{cases} p_u - \alpha_u & \text{if } u \notin \bigcup_{p=0}^{P} N_p, \\ V_p(u) & \text{if } u \in N_p. \end{cases}$$

This leads to the following easily verified result.

LEMMA 6.2. *The Lyapunov function V is continuous. Furthermore,*

(i) *if $\widetilde{\varphi}([0,t], u) \bigcap (\bigcup_{p=0}^{P} N_p) = \emptyset$, then*

$$V(\widetilde{\varphi}(t, u)) = V(u) - t;$$

(ii) *if $\widetilde{\varphi}([0,t], u) \subset N_p$, $\widetilde{\varphi}([0,t], u') \subset N_p$, $\tau_p(u) = \tau_p(u')$, and $V(u) = V(u')$, then*

$$V(\widetilde{\varphi}(t, u)) = V(\widetilde{\varphi}(t, u')).$$

We now define

$$\begin{aligned} \hat{f}(u) &= (V(u), \tau(u)) \\ &= (V(u), \tau_0(u), \tau_1(u), \ldots, \tau_{P-1}(u)) \end{aligned}$$

By this definition it is clear that $\hat{f} : \mathcal{A} \to [0, P] \times [-1, 1]^P$. The following lemma implies that, for our purposes, we can consider the range of $\hat{f}$ to be $\partial\left([0, P] \times [-1, 1]^P\right)$.

LEMMA 6.3. $\hat{f}(\mathcal{A}) \subset \partial\left([0, P] \times [-1, 1]^P\right)$, *and hence*

$$\hat{f} : \mathcal{A} \to \partial\left([0, P] \times [-1, 1]^P\right).$$

Proof. Let $u \in \mathcal{A}$. Then $\alpha(u) \subset M(p^*)$ and $\omega(u) \subset M(q^*)$ for some $0 \le q \le p \le P$. If $u = M(P)$ then

$$\widetilde{f}(u) = (P, 0, 0, \ldots, 0) \in \partial([0, P] \times [-1, 1]^P).$$

Therefore, assume that $u \neq M(P)$; then $\omega(u) \in M(p)$, where $p < P$, and hence $\tau_p(u) = \pm 1$. □

6.2. $\hat{f} = \mathcal{Q} \circ \tilde{f}$ **is continuous.** We shall now define the equivalence relation on $\partial\left([0,P] \times [-1,1]^P\right)$. We shall consider the space $\partial\left([0,P] \times [-1,1]^P\right) \times \{1\}$ for notational convenience and denote elements of this space by

$$(v,\tau) = (v,\tau_0,\tau_1,\ldots,\tau_P),$$

where $\tau_P \equiv 1$.

For $q,p \in \{0,1,\ldots,P\}$, $w \in [0,P]$, and $q \le w \le p$ let

$$B(q,w,p) = \left\{ (v,\tau) \in \partial([0,P] \times [-1,1]^P) \times \{1\} \;\middle|\; \begin{array}{l} v = w, \\ \tau_P := 1, \\ \tau_p = \pm 1, \\ \tau_q = \pm 1, \\ -1 < \tau_r < 1 \text{ for } q < r < p \end{array} \right\}.$$

If $(v,\tau) \in \partial([0,P] \times [-1,1]^P) \times \{1\}$ then there exist q,p such that $(v,\tau) \in B(q,v,p)$. For $(v,\tau) \in B(q,v,p)$ define

$$(v,\tau) \sim (v,\tau^*),$$

where $\tau_l^* = 0$ if $l < q$ or $l > p$ and $\tau_r^* = \tau_r$ for $q \le r \le p$.

Now let

$$\mathcal{Q} : \partial\left([0,P] \times [-1,1]^P\right) \to \partial\left([0,P] \times [-1,1]^P\right) / \sim$$

be the quotient map induced by this equivalence relation.

PROPOSITION 6.4. $\tilde{f} : \mathcal{A} \to D^P$ *is continuous.*

Proof. Since $\mathcal{Q}$ and V are continuous, it is clear that any possible lack of continuity of $\hat{f}$ must arise from the map τ. As shall be shown, the discontinuities induced by τ are eliminated via the quotient map $\mathcal{Q}$. Recall that τ_p is continuous on Θ_p. Thus, when checking for discontinuities induced by τ_p, one need only consider $u \in \mathcal{A} \setminus \Theta_p$.

With this in mind, consider $u \in \mathcal{A}$ such that $q < V(u) < p$, $\tau_q(u) = \pm 1$, and $\tau_p(u) = \pm 1$. This implies that $\omega(u) \subset M(q^{\pm})$ and $\alpha(u) \subset M(p^{\pm})$. In addition, for $r < q$ or $r > p$, $\tau_r(u) = 0$ and for $q < r < p$, $-1 < \tau_r(u) < 1$. Let $\{u_n\} \subset \mathcal{A}$ such that $u_n \to u$ as $n \to \infty$. By continuity of the flow, for n sufficiently large and $q < r < p$, $I_r(u)$ is a uniformly bounded (possibly empty) set. Thus $\tau_r(u_n) \to \tau_r(u)$ as $n \to \infty$. Obviously, if $\tau_r(u_n) \to \tau_r(u)$ for all r then we are done. Therefore, without loss of generality, it may be assumed that, for some fixed $r < q$ or $r > p$, $\tau_r(u_n) = 1$. Let us assume that $r > p$. This and the continuity of V imply that $\hat{f}(u_n) \to (V(u),\bar{\tau})$, where

$$(V(u),\bar{\tau}) = (V(u),\bar{\tau}_0,\ldots,\bar{\tau}_{q-1},1,\bar{\tau}_{q+1},\ldots,\bar{\tau}_{p-1},1,\bar{\tau}_{p+1},\ldots,\bar{\tau}_{r-1},1,\bar{\tau}_{r+1},\ldots,\bar{\tau}_{P-1}).$$

But by definition this makes

$$(V(u),\bar{\tau}) \sim (V(u),0,\ldots,0,1,\bar{\tau}_{q+1},\ldots,\bar{\tau}_{p-1},1,0,\ldots,0) = \hat{f}(u).$$

Therefore, $\tilde{f}$ is continuous. □

6.3. $\tilde{f} : \mathcal{A} \to D^P$**.** It can be easily checked that $\partial\left([0,P] \times [-1,1]^P\right) / \sim$ is not homeomorphic to D^P. Therefore, it needs to be shown that

$$\hat{f}(\mathcal{A})/\sim \subset Y \subset \partial\left([0,P] \times [-1,1]^P\right) / \sim,$$

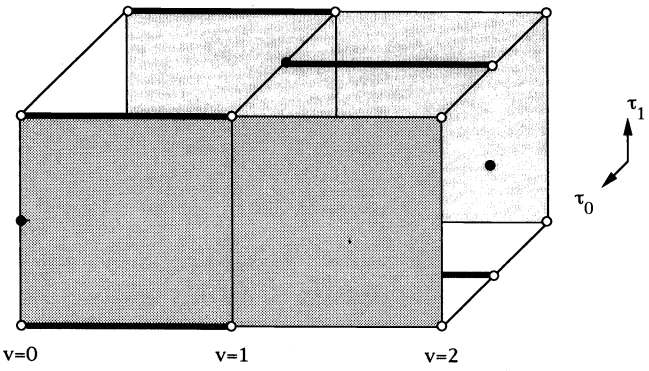


FIG. 4. *The shaded regions and dark lines indicate* $\Xi := \bigcup_{w=0}^{P} \Xi_w$ *for* $P = 3$.

where Y is homeomorphic to D^P. We begin by sharpening our understanding of $\hat{f}(\mathcal{A}) \subset \partial\left([0,P] \times [-1,1]^P\right)$. Let $F_s = \{(v, \tau_0, \ldots, \tau_{P-1}) \mid \tau_s = \pm 1\}$.

LEMMA 6.5. *If* $V(u) \in [q, q+1)$ *then, for all* $s > q$,

$$\operatorname{int}(F_s) \cap f(\mathcal{A}) = \emptyset.$$

Proof. Let $f(u) \in F_s$, where $s > q$. Now $V(u) \in [q, q+1)$ implies that $\omega(u) \subset M(r)$ for some $r \leq q$. Thus $f(u) \in F_s \cap F_r \subset \partial F_s$. □

Let $w \in [q, q+1)$ and define

$$\Xi_w = \bigcup_{s=0}^{q} F_s.$$

It is easy to see that Ξ_w is homeomorphic to $S^q \times [-1,1]^{P-q-1}$.

LEMMA 6.6. *Let* $\mathcal{A}_w = \{u \in \mathcal{A} \mid V(u) = w\}$. *Then*

$$\hat{f}(\mathcal{A}_w) \subset \Xi_w.$$

Proof. Let $w \in [q, q+1)$. $V(u) = w$ implies that $\omega(u) \subset M(r)$, where $r \leq q$. Thus, $\hat{f}(u) \subset F_r \subset \Xi_w$. □

Since $\hat{f}(\mathcal{A}_w) \subset \Xi_w$, $\hat{f}(\mathcal{A}) \subset \Xi := \bigcup_{w=0}^{P} \Xi_w$. We shall show that $\mathcal{Q}(\Xi)$ is homeomorphic to D^p. To do this we return to the flow ψ defined in the introduction and remark that it satisfies assumptions (A1)–(A4) with $\mathcal{A} = D^p$. In particular, we shall let

$$\mathcal{M}(D^P, \psi) := \{\Pi(p^{\pm}) \mid \Pi(p^{\pm}) = \mathbf{e}_p^{\pm}\} \cup \{\Pi(P) = 0\}$$

denote the Morse decomposition for D^P under ψ. Observe that the flow-defined order is

$$P > P - 1^{\pm} > \cdots > 1^{\pm} > 0^{\pm}.$$

The point of this observation is that we can now apply the results of the previous sections to this flow. Therefore, there exists a function $\hat{g} : D^P \to \partial\left([0,P] \times [-1,1]^P\right)$ given by $\hat{g}(z) = (V(z), \tau(z))$ defined as in §5.1. By §5.2 and Lemma 6.4,

$$\tilde{g} := \mathcal{Q} \circ \hat{g} : D^P \to \mathcal{Q}(\Xi)$$

is continuous.

From the construction of $\hat{g}$ it is clear that this function is dependent upon the choice of isolating neighborhoods $N_{p\pm}$.

PROPOSITION 6.7. *There exists a choice of isolating neighborhoods $N_{p\pm}$, satisfying Proposition 4.4 such that*

$$\tilde{g} : D^P \to \mathcal{Q}(\Xi)$$

is a homeomorphism.

Proof. We leave it to the reader to check that $\mathcal{Q}(\Xi)$ is Hausdorff. Obviously, D^P is compact and, by Lemma 6.4, $\tilde{g}$ is continuous. Thus it is sufficient to find isolating neighborhoods $N_{p\pm}$ such that $\tilde{g}$ is a bijection.

Let

$$K_\delta(p) := \sum_{i \neq p} |z_i| - |z_p| \leq \delta$$

for some $0 < \delta << 1$. Let $B_\delta(0) = \{z \in \mathbf{R}^P \mid \|z\| \leq \delta\}$. Define

$$N_P = B_{\frac{\delta}{2}}(0),$$

and for $p = 0, \ldots, P-1$ set

$$N_p = \left(K_{(P-p)\delta}(p) \setminus \left(\bigcup_{q>p} K_{(P-p)\delta}(q) \cup B_{(P-p)\delta}(0) \right) \right) \cap D^P.$$

Finally, let $N_{p\pm}$ denote the component of N_p which contains $\mathbf{e}_p^\pm$, respectively.

We leave it to the reader to check that, using these neighborhoods, $\hat{g}$ is 1-1. The laborious part is showing that, given p, the set

$$\left\{ z \in D^P \;\middle|\; \begin{array}{l} \omega(z) = \mathbf{e}_q^\pm \text{ for some } q \leq p \\ \alpha(z) = \mathbf{e}_r^\pm \text{ for some } r \geq p \end{array} \right\}$$

is parameterized by

$$\left\{ z \in D^P \;\middle|\; \sum_{i \neq p} |z_i| - |z_p| = (P-p)\delta \right\}.$$

Observe that on $\hat{g}(D^P)$, $\mathcal{Q}$ is 1-1. This implies that $\tilde{g}$ is 1-1.

Finally, recall that, by definition, given $[(v,\tau)] \in \partial\left([0,P] \times [-1,1]^P\right) / \sim$, there exists τ^* of the form $\tau_q^* = \pm 1$, $\tau_p^* = \pm 1$, $-1 < \tau_r < 1$ for $q < r < p$, and $\tau_r = 0$ if $r < q$ or $r > p$ such that

$$(v,\tau) \sim (v,\tau^*).$$

But $(v,\tau^*) \in \tilde{g}(D^P)$. In particular, $\omega(\tilde{g}^{-1}(v,\tau^*)) = \Pi(q^s)$, where $s = \operatorname{sgn} \tau_q$ and $\alpha(\tilde{g}^{-1}(v,\tau^*)) = \Pi(p^s)$, where $s = \operatorname{sgn} \tau_p$. Furthermore, for $q < r < p$, τ_r denotes how "close" the trajectory of $\tilde{g}^{-1}((v,\tau^*)$ passes by $\Pi(r^*)$. Since we know the flow on D^P, we can check that such an orbit exists for each value of τ_r. Therefore, $\tilde{g}$ is a bijection. □

6.4. The induced flow $\widetilde{\psi}$. To derive the existence of the induced flow $\widetilde{\psi}$ we shall prove the existence of the following commutative diagram:

$$\begin{array}{ccccc} \mathbf{R}\times\mathcal{A} & \overset{id\times\hat{f}}{\rightarrow} & \mathbf{R}\times\Xi & \overset{id\times\mathcal{Q}}{\rightarrow} & \mathbf{R}\times D^P \\ \downarrow\widetilde{\varphi} & & \downarrow\hat{\psi} & & \downarrow\widetilde{\psi} \\ \mathcal{A} & \overset{\hat{f}}{\rightarrow} & \Xi & \overset{\mathcal{Q}}{\rightarrow} & D^P \end{array} \tag{18}$$

LEMMA 6.8. *$\widehat{\psi}$ exists.*

Proof. Define

$$\widehat{\psi}(t,\widehat{f}(u)) = (V(\widetilde{\varphi}(t,u)),\tau(u))\,.$$

The first step is to check that this expression is well defined. Observe that $\tau(u) = \tau(\widetilde{\varphi}(t,u))$ for all $t\in\mathbf{R}$. Furthermore, recall that if $u,u'\in\mathcal{A}$ such that $\tau(u)=\tau(u')$ and $V(u)=V(u')$, then $V(\widetilde{\varphi}(t,u))=V(\widetilde{\varphi}(t,u'))$ for all $t\in\mathbf{R}$.

The left square in diagram (18) commutes since

$$\begin{aligned}\hat{\psi}((id\times\hat{f})(t,u)) &= \hat{\psi}(t,\hat{f}(u)) \\ &= (V(\widetilde{\varphi}(t,u)),\tau(u)) \\ &= (V(\widetilde{\varphi}(t,u)),\tau(\widetilde{\varphi}(t,u))) \\ &= \hat{f}(\widetilde{\varphi}(t,u)).\end{aligned}$$

At this point $\hat{\psi}$ is defined only on $\hat{f}(\mathcal{A})$, not on all of Ξ. However, if we recall the properties of $\widetilde{\varphi}$, τ, and V, then we can make the following observations:

1. The flow lines on $\partial\left([0,P]\times[-1,1]^P\right)$ are "horizontal" lines, i.e., lines parallel to the v axis.
2. On the sets $\{(v,\tau)\mid p+\frac{1}{4}\le v\le p+\frac{3}{4}\}$, where $p=0,\dots,P-1$, the flow $\hat{\psi}$ is simply uniform translation in the $-v$ direction.
3. On the sets $\{(v,\tau)\mid p-\frac{1}{4}\le v\le p+\frac{1}{4}\}$ the velocity of the flow $\hat{\psi}$ depends smoothly on τ_p.

Therefore, $\hat{\psi}$ can be extended to all of Ξ in such a way that it respects the quotient map. This, in turn, implies that

$$\widetilde{\psi}(t,\mathcal{Q}(v,\tau)) := \mathcal{Q}(\hat{\psi}(t,(v,\tau)))$$

is well defined, and hence the right square of (18) also commutes. □

From the remarks in this proof it is, we hope, clear that the flow $\hat{\psi}$ on Ξ is, in fact, independent of the maps $\widetilde{\varphi}$, τ, and V. This, in turn, implies that the flow $\widetilde{\psi}$ on D^P is predetermined. Therefore, the conjugacy $\widetilde{g}$ of the previous section relates the flow ψ with $\widetilde{\psi}$. In other words, if we define

$$f = \widetilde{g}^{-1}\circ\widetilde{f},$$

then $f:\mathcal{A}\to D^P$ is the semiconjugacy of Theorem 1.2.

7. Surjectivity of f. In the previous section we constructed a proper semiconjugacy $f:\mathcal{A}\to D^P$ between the flow $\widetilde{\varphi}$ and ψ. Now we shall show that f is, in fact, surjective. The basis of our result is the following elementary fact from algebraic topology.

LEMMA 7.1. *If $f : X \to S^n$ is continuous and $f^* : \check{H}^n(S^n) \to \check{H}^n(X)$ is nonzero, then f is surjective.*

Our strategy is to reduce the cohomology index information to an application of this lemma.

Recall that $\{\Pi(p^{\pm}) \mid p = 0, \dots, P-1\} \cup \{\Pi(P)\}$ with the flow-defined partial order $P > P-1^{\pm} > \cdots > 0^{\pm}$ is a Morse decomposition for D^P under ψ. Thus the indexing set and flow-defined order for $\mathcal{M}(D^P)$ and $\mathcal{M}(\mathcal{A})$ are identical. This implies that I is an interval in $\mathcal{M}(D^P)$ if and only if it is an interval for $\mathcal{M}(\mathcal{A})$. For an interval I define

$$f_I := f|_{M(I)} : M(I) \to D^P.$$

This of course implies that $f_I(M(I)) \subset \Pi(I)$.

Since both $\mathcal{A}$ and D^P are attractors, [20, Prop. 7.4] implies that, for any interval I,

$$f^* : CH^*(\Pi(I)) \to CH^*(M(I))$$

is an isomorphism. From now on let

$$I = \{0^{\pm}, \dots, P-1^{\pm}\}.$$

I is an attracting interval, and hence $M(I)$ and $\Pi(I)$ are attractors. Therefore, the maps $\iota_I^0 : \check{H}^k(\Pi(I)) \to CH^*(\Pi(I))$ and $\iota_I^{0'} : \check{H}^k(M(I)) \to CH^*(M(I))$ are isomorphisms.

By [20, Prop. 7.4] there exists a commutative diagram

$$\begin{array}{ccc} \check{H}^k(\Pi(I)) & \xrightarrow{\iota_I^0} & CH^k(\Pi(I)) \\ \downarrow f_I^* & & \downarrow f^* \\ \check{H}^k(M(I)) & \xrightarrow{\iota_I^{0'}} & CH^k(M(I)). \end{array} \tag{19}$$

This now forces f_I^* to be an isomorphism. Since $\Pi(I) = S^{P-1}$, Lemma 7.1 implies that $f : M(I) \to \Pi(I)$ is surjective.

Since $f(M(P)) = \Pi(P)$, it only remains to be shown that $C(M(P), M(I))$ maps onto $C(\Pi(P), \Pi(P))$. To do this we use the fact that

$$\mathcal{W} := \left\{ u \in \mathcal{A} \mid V(u) = P - \frac{1}{2} \right\}$$

is a local section for $\widetilde{\varphi}$ and

$$W := \tilde{g}^{-1}(\mathcal{Q}(\Xi_{P-\frac{1}{2}}))$$

is a local section for ψ. Since $\widetilde{\varphi}(\mathbf{R}, W) = C(M(P), M(I))$ and $\psi(\mathbf{R}, \mathcal{W}) = C(\Pi(P), \Pi(I))$, it is sufficient to show that $f : \mathcal{W} \to W$ is onto. Observe that $\mathcal{Q}(\Xi_{P-\frac{1}{2}})$, and hence W is homeomorphic to S^{P-1}.

Define

$$N := \widetilde{g}^{-1}\left(\mathcal{Q}\left(\bigcup_{w=P-\frac{1}{2}}^{P} \Xi_w\right)\right).$$

Then, (N, W) is an index pair for $\Pi(P)$. Furthermore, $(\mathcal{N}, \mathcal{W}) := (f^{-1}(N), \mathcal{W})$ is an index pair for $M(P)$. We can relate this information via the following commutative diagram:

$$\begin{array}{ccccccccc} \to & \check{H}^{P-1}(N) & \to & \check{H}^{P-1}(W) & \stackrel{\delta}{\to} & \check{H}^{P}(N,W) & \to & \check{H}^{P}(N) & \to \\ & \downarrow f^* & & \downarrow f^* & & \downarrow f^* & & \downarrow f^* & \\ \to & \check{H}^{P-1}(\mathcal{N}) & \to & \check{H}^{P-1}(\mathcal{W}) & \stackrel{\delta'}{\to} & \check{H}^{P}(\mathcal{N},\mathcal{W}) & \to & \check{H}^{P}(\mathcal{N}) & \to . \end{array}$$

Using the fact that N is homeomorphic to D^P and W is homeomorphic to S^{P-1}, and using the cohomology index information, this reduces to

$$\begin{array}{ccccccccc} \to & 0 & \to & \mathbf{Z} & \stackrel{\delta}{\to} & \mathbf{Z} & \to & 0 & \to \\ & \downarrow & & \downarrow f^* & & \downarrow \approx & & \downarrow & \\ \to & \check{H}^{P-1}(\mathcal{N}) & \to & H^{P-1}(\mathcal{W}) & \stackrel{\delta'}{\to} & \check{H}^{P}(\mathcal{N},\mathcal{W}) & \to & H^{P}(\mathcal{N}) & \to . \end{array}$$

Therefore, $f^* : \check{H}^{P-1}(W) \to \check{H}^{P-1}(\mathcal{W})$ is injective. By Lemma 7.1 this implies that $f : \mathcal{W} \to W$ is onto.

Acknowledgments. My original interest in this problem was motivated by conversations with H. Hattori. The idea of using semiconjugacies to describe the global dynamics grew out of discussions with C. McCord. Finally, I would like to thank C. Grant for explanations concerning the Cahn–Hilliard and phase-field equations.

REFERENCES

[1] S. Angenant, *The Morse–Smale property for a semilinear parabolic equation*, J. Differential Equations, 62 (1986), pp. 427–442.

[2] P. Bates and P. Fife, *Spectral comparison principles for the Cahn–Hilliard and phase-field equations, and times scales for coarsening*, Phys. D, 43 (1990), pp. 335–348.

[3] P. Bates and S. Zheng, *Inertial manifolds and inertial sets for the phase-field equations*, J. Dyn. Diff. Eqs., 4 (1992), pp. 375–398.

[4] P. Brunovsky and B. Fiedler, *Connecting orbits in scalar reaction diffusion equations*, in Dynam. Report., 1, U. Kirchgraber and H.O. Walter, eds., 1988, pp. 57–89.

[5] N. Chafee and E. Infante, *A bifurcation problem for a nonlinear parabolic equation*, J. Appl. Anal., 4 (1974), pp. 17–37.

[6] C. Conley, *Isolated Invariant Sets and the Morse Index*, in CBMS Lecture Notes 38, American Mathematical Society, Providence, RI, 1978.

[7] C. Conley and J. Smoller, *Bifurcation and stability of stationary solutions of Fitz–Hugh–Nagumo equations*, J. Differential Equations, 63 (1986), pp. 389–405.

[8] P. Fife, *Models for phase separation and their mathematics*, in Nonlinear Partial Differential Equations and Applications, M. Mimura and T. Nishida, eds., Kinokuniya Publishers, Tokyo, Japan, to appear.

[9] R. Franzosa, *The connection matrix theory for Morse decompositions*, Trans. Amer. Math. Soc., 311 (1989), pp. 561–592.

[10] R. Franzosa and K. Mischaikow, *Algebraic transition matrices*, Georgia Institute of Technology, Atlanta, GA, CDSNS94-203, preprint.

[11] J. K. Hale, *Asymptotic Behaviour of Dissipative Systems*, in Math. Surveys Monographs 25, American Mathematical Society, Providence, RI, 1988.

[12] J. K. Hale, L. T. Magalhães, and W. M. Oliva, *An introduction to infinite dimensional dynamical systems—geometric theory*, in Appl. Math. Sci. 47, Springer-Verlag, New York, 1984.

[13] J. K. Hale and G. Raugel, *Partial differential equations on thin domains*, Proc. 1990 UAB Internat. Conference. Differential Equations Math. Phys., Academic Press, New York, to appear.

[14] H. Hattori and K. Mischaikow, *A dynamical system approach to a phase transition problem*, J. Differential Equations, 94 (1991), pp. 340–378.

[15] D. HENRY, *Geometric theory of semilinear parabolic equations*, in Springer Lecture Notes 840, Springer–Verlag, New York, Berlin, Heidelberg, 1981.
[16] ———, *Some infinite dimensional Morse–Smale systems defined by parabolic differential equations*, J. Differential Equations, 59 (1985), pp. 165–205.
[17] C. MCCORD, *Mappings and homological properties in the homology Conley index*, Engrg. Theory and Dynam. Systems, 8* (1988), pp. 175–198.
[18] ———, *Mappings and Morse decompositions in the homology Conley index*, Indiana Univ. Math. J., 40 (1991), pp. 1061–1082.
[19] ———, *The connection map for attractor repeller pairs*, Trans. Amer. Math. Soc., 307 (1988), pp. 195–203.
[20] C. MCCORD AND K. MISCHAIKOW, *On the global dynamics of attractors for scalar delay equations*, J. Amer. Math. Soc., to appear, CDSNS92-89, preprint.
[21] K. P. RYBAKOWSKI, *The Homotopy Index and Partial Differential Equations*, in Universitext, Springer-Verlag, New York, Berlin, Heidelberg, 1987.
[22] D. SALAMON, *Connected simple systems and the Conley index of isolated invariant sets*, Trans. Amer. Math. Soc., 291 (1985), pp. 1–41.
[23] J. SMOLLER, *Shock Waves and Reaction Diffusion Equations*, Springer-Verlag, New York, 1983.
[24] R. TEMAM, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1988.
[25] S. ZHENG, *Asymptotic behavior of solutions to the Cahn–Hilliard equation*, Appl. Anal., 23 (1986), pp. 165–184.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1225–1232, September 1995

008

THE EXISTENCE OF PERIODIC SOLUTIONS TO REACTION-DIFFUSION SYSTEMS WITH PERIODIC DATA *

J. J. MORGAN† AND S. L. HOLLIS‡

Abstract. The existence of time-periodic solutions is proven for a large class of reaction-diffusion systems in which Dirichlet boundary data, diffusivities, and reaction rates are periodic with common period.

Key words. periodic solutions, reaction-diffusion systems

AMS subject classifications. 35B10, 35K45, 35K57

1. Introduction.

We consider reaction-diffusion systems of the form

$$\begin{aligned} \frac{\partial u_i}{\partial t} - d_i(t)\Delta u_i &= f_i(x,t,u) && \text{in } \Omega\times\{t>0\},\ i=1,\dots,m, \\ u_i(x,t) &= g_i(x,t) && \text{on } \partial\Omega\times\{t>0\},\ i=1,\dots,m, \\ u_i(x,0) &= u_{0_i}(x) && \text{on } \overline{\Omega},\ i=1,\dots,m, \end{aligned} \tag{1.1}$$

where $u = (u_i)_{i=1}^m$ and Ω is a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$; i.e., $\partial\Omega$ is an $(n-1)$-dimensional $C^{2+\alpha}$ manifold of which Ω lies locally on one side. We assume that the initial data u_{0_i} are bounded, measurable, and nonnegative and each $d_i \in C(\mathbb{R}_+;[a,b])$, where $0 < a \le b < \infty$. (The symbol $\mathbb{R}_+$ denotes $[0,\infty)$.) We also assume that the reaction functions f_i are continuous on $\overline{\Omega}\times\mathbb{R}_+\times\mathbb{R}_+^m$ and locally Lipschitz in u, and $f = (f_i)_{i=1}^m$ is *quasi positive*; i.e., for each $i = 1,\dots,m$, we have $f_i(\cdot,\cdot,\xi) \ge 0$ for all $\xi \ge 0$ with $\xi_i = 0$. Each g_i is assumed to be a nonnegative member of $C^{2,1}(\partial\Omega\times\mathbb{R}_+)$. These standard basic assumptions guarantee *local* existence of unique, nonnegative, classical solutions on a maximal time interval $0 \le t < T^* \le \infty$. This follows from a straightforward adaptation of results in, e.g., [5] and [13] to account for the t dependence of the parameters in (1.1).

In addition to the basic assumptions stated above, we assume the following:

(A1) There is a $K > 0$, and for each $i = 1,\dots,m$ there are nonnegative constants $\alpha_{i,1},\alpha_{i,2},\dots,\alpha_{i,i}$ with $\alpha_{i,i} > 0$ such that

$$\sum_{j=1}^{i}\alpha_{i,j}f_j(x,t,\xi) \le K\left(1+\sum_{j=1}^{m}\xi_j\right) \text{ for all } x\in\Omega,\ t\ge 0, \text{ and } \xi\in\mathbb{R}_+^m.$$

(A2) Each $|f_i(\cdot,\cdot,\xi)|$, $i = 1,\dots,m$ is bounded above by a polynomial in $\xi_1,\dots,\xi_m$.

The following global existence theorem follows from results in Morgan [10].

THEOREM 1.1. *Let conditions* (A1) *and* (A2) *be met. Then, for any bounded, measurable, nonnegative initial data* $u_0 = (u_{0_i})_{i=1}^m$, *we have* $T^* = \infty$; *i.e., system* (1.1) *has a unique, nonnegative, classical solution on* $\overline{\Omega}\times[0,\infty)$.

*Received by the editors September 25, 1993; accepted for publication January 17, 1994.

†Department of Mathematics, Texas A&M University, College Station, Texas 77843. The research of this author was supported in part by National Science Foundation grant DMS-9208046.

‡Department of Mathematics and Computer Science, Armstrong State College, Savannah, Georgia 31419.

Remark. We refer to (A1) as a linear "intermediate sums" condition [10]. It allows high-order nonlinearities in the individual f_i but requires that f_1 be bounded above by a linear polynomial in u, and cancellation of high-order positive terms in the intermediate sums. An illustrative example is the following three-species predator-prey system:

$$\begin{aligned} \frac{\partial u_1}{\partial t} - d_1 \Delta u_1 &= k_{11} u_1 (M - u_1) - k_{12} u_1 u_2 - k_{13} u_1 u_3, \\ \frac{\partial u_2}{\partial t} - d_2 \Delta u_2 &= k_{12} u_1 u_2 - k_{23} u_2 u_3 - k_{20} u_2, \\ \frac{\partial u_3}{\partial t} - d_3 \Delta u_3 &= k_{13} u_1 u_3 + k_{23} u_2 u_3 - k_{30} u_3, \end{aligned} \tag{1.2}$$

where M and the k_{ij} are bounded, nonnegative, continuous functions on $\overline{\Omega} \times \mathbf{R}_+$. Here one can take $\alpha_{2,1} = \alpha_{2,2} = 1$ and $\alpha_{3,1} = \alpha_{3,2} = \alpha_{3,3} = 1$. Note that in this example we actually have $\sum_{j=1}^{i} \alpha_{i,j} f_j(x,t,u) \leq K$ because of the assumption of logistic growth of u_1 when $u_2 = u_3 = 0$. Exponential growth of one or more of the species in the absence of the others would lead to a condition of precisely the type in (A1).

An intermediate sums condition of the form (A1) is indeed satisfied by a variety of complex models of, e.g., population dynamics, chemical reactions, and spread of disease [10]–[12]. We also remark that nonlinear intermediate sums are possible with the allowable order depending upon the spatial dimension n; see [10].

Our concern here is the existence of a time-periodic solution to (1.1) in the situation where the reaction function f, the diffusivities d_i, and the boundary data g are each periodic in t with common period T. Of particular interest from the point of view of applications would be, e.g.,

- population models and models of the spread of disease in which birth and death rates, rates of diffusion, rates of infection/interaction, and environmental carrying capacities are periodic on a seasonal scale;
- chemical reaction models in which reaction rates and diffusivities are periodic on a daily scale because of oscillations in sunlight and/or temperature.

With this issue in mind, we assume the following:

(A3) There is a $T > 0$ such that, for $i = 1, \ldots, m$ and $t \geq 0$, we have

$$f_i(\cdot, t, \cdot) = f_i(\cdot, t+T, \cdot), \quad g_i(\cdot, t) = g_i(\cdot, t+T), \quad \text{and} \quad d_i(t) = d_i(t+T).$$

(A4) There is a continuous function $\tilde{g} : \overline{\Omega} \to \mathbf{R}_+^m$ such that $g(\cdot, 0) = g(\cdot, T) = \tilde{g}|_{\partial\Omega}$.

(A5) The constants K and $\alpha_{m,1}, \ldots, \alpha_{m,m}$ in (A1) may be chosen so that $\alpha_{m,j} > 0$ for $j = 1, \ldots, m$ and

$$\sum_{j=1}^{m} \alpha_{m,j} f_j(x,t,\xi) \leq K \quad \text{for all } x \in \Omega,\ t \geq 0, \text{ and } \xi \in \mathbf{R}_+^m.$$

Note that (A5) is satisfied by the example system (1.2). This would also be true of more general population models of this type provided that each species exhibits bounded growth in the absence of all other species.

Our main result is the following theorem.

THEOREM 1.2. *Under assumptions* (A1)–(A5) *there exists a* $u_0 \in C(\overline{\Omega}; \mathbf{R}_+^m)$ *such that the solution of* (1.1) *satisfies* $u(\cdot, t) = u(\cdot, t+T)$ *on* $\overline{\Omega}$ *for all* $t \geq 0$.

Previous results along these lines can be found in Liu and Pao [9], where the existence of a (unique) T-periodic solution is established via the contraction mapping theorem in the case of a one-dimensional domain and under somewhat stringent conditions on the diffusion coefficients and reaction rates. Our approach will use a variation on Schauder's theorem and will require no assumptions other than (A1)–(A5) to establish the existence of a T-periodic solution. Related work in which scalar parabolic equations are considered includes [1]–[3] and [14].

2. Formulation of the fixed point problem. We will use the following corollary to Schauder's theorem. For the proof see, e.g., Gilbarg and Trudinger [4, Thm. 11.3].

THEOREM 2.1. *Let X be a Banach space and $F : X \to X$ be a compact map. Assume that there exists a constant $C > 0$ such that $\|z\| < C$ for all z satisfying $z = \sigma Fz$ with $\sigma \in (0,1)$. Then there exists a fixed point z^* of F satisfying $\|z^*\| \leq C$.*

For convenience of notation, let us define the formal solution operator for (1.1) by $\mathcal{S}(t)u_0 = u(\cdot\,, t)$ for $t \geq 0$. Now define $F : C_0(\overline{\Omega}; \mathbf{R}^m) \to C_0(\overline{\Omega}; \mathbf{R}^m)$ by

$$Fz = \mathcal{S}(T)(z + \tilde{g})^+ - \tilde{g}, \tag{2.1}$$

where

$$C_0(\overline{\Omega}; \mathbf{R}^m) = \{z \in C(\overline{\Omega}; \mathbf{R}^m) \mid z = 0 \text{ on } \partial\Omega\}$$

and T and $\tilde{g}$ are as in (A3) and (A4). By parabolic regularity [8], F is a compact map. Note also that if z^* is a fixed point of F, then $z^* + \tilde{g} = \mathcal{S}(T)(z^* + \tilde{g})^+$. Consequently, $z^* + \tilde{g} \geq 0$, so $u_0^* \equiv z^* + \tilde{g}$ is a (nonnegative) fixed point of $\mathcal{S}(T)$. So the existence of a T-perodic solution of (1.1) will follow from the existence of a fixed point of the operator F because of (A3) and uniqueness of solutions to (1.1).

Suppose that $0 < \sigma < 1$ and $z = \sigma Fz$. Also, set $u_0 = z + \tilde{g}$. Then we see that

$$u_0 = \sigma \mathcal{S}(T)u_0^+ + (1 - \sigma)\tilde{g}.$$

But $z = \sigma Fz$ implies that $z + \sigma\tilde{g} \geq 0$, which then implies that $u_0 \geq 0$. Thus

$$u_0 = \sigma \mathcal{S}(T)u_0 + (1 - \sigma)\tilde{g}. \tag{2.2}$$

Let us now define the set

$$\Lambda_T = \{u_0 \in C(\overline{\Omega}; \mathbf{R}_+^m) \mid u_0 = \sigma \mathcal{S}(T)u_0 + (1 - \sigma)\tilde{g} \text{ for some } \sigma \in (0,1)\}. \tag{2.3}$$

In light of Theorem 2.1 and the preceding observations, our goal is to show that Λ_T is a bounded subset of $C(\overline{\Omega}; \mathbf{R}_+^m)$. Note that, because of (2.2), this can be accomplished by showing that there is a $C > 0$ such that $\|\mathcal{S}(T)u_0\|_\infty \leq C$ for all $u_0 \in \Lambda_T$.

3. A preliminary estimate. Our first step toward showing that Λ_T is a bounded subset of $C(\overline{\Omega}; \mathbf{R}_+^m)$ is the following L^1 estimate.

LEMMA 3.1. *Suppose that* (A1)–(A5) *are true. Then there is a constant $C_1 > 0$ such that*

$$\|u_i\|_{1,\Omega\times(0,T)} \leq C_1, \quad i = 1, \ldots, m$$

for all u satisfying (1.1) *with $u_0 \in \Lambda_T$.*

Proof. Let $u_0 \in \Lambda_T$ and u be the corresponding solution of (1.1). Also, define $w \equiv \int_0^T \sum_{k=1}^m \alpha_{m,k} d_k(s) u_k(\cdot, s) ds$. Summing the equations in (1.1), applying (A5), and integrating over $t \in [0,T]$ yields

$$\sum_{k=1}^m \alpha_{m,k}\big(u_k(\cdot,T) - u_{0_k}\big) - \Delta w \le KT \quad \text{on } \Omega. \tag{3.1}$$

For convenience, set $v = F(u_0 - \tilde{g})$, where F is as in (2.1). Then $u(\cdot,T) = v + \tilde{g}$ and $u_0 = \sigma v + \tilde{g}$. So (3.1) becomes

$$\sum_{k=1}^m \alpha_{m,k}(v_k + \tilde{g}_k) - \Delta w \le \sum_{k=1}^m \alpha_{m,k}(\sigma v_k + \tilde{g}_k) + KT \quad \text{on } \Omega.$$

Hence

$$-\Delta w \le (\sigma - 1)\sum_{k=1}^m \alpha_{m,k} v_k + KT \le (1-\sigma)\sum_{k=1}^m \alpha_{m,k}\tilde{g}_k + KT \quad \text{on } \Omega$$

since $v_k \ge -\tilde{g}_k$. Also, on $\partial\Omega$ we have $w = \int_0^T \sum_{k=1}^m \alpha_{m,k} d_k(s) g_k(\cdot, s) ds$. Therefore, one can apply a comparison principle and nonnegativity to obtain a bound on $\|w\|_{\infty,\Omega}$ and, in turn, a bound on $\|\sum_{k=1}^m u_k\|_{1,\Omega\times(0,T)}$, where each bound is independent of u_0 and σ. □

Remark. This result remains true without (A1) and (A2) provided that the interval $[0,T]$ lies within the maximal interval of existence $[0,T^*)$.

4. The bootstrapping framework. The following lemma provides a bootstrapping mechanism for obtaining L^p estimates for large p from an L^1 estimate. Although the proof is essentially the same as that of similar results in [6], [7], [10], and [11], we include it here for the sake of completeness.

LEMMA 4.1. *Suppose that* (A1) *is true and* u *satisfies* (1.1) *for* $0 \le t < T$. *Let* $\tau \in [0,T)$. *There is a constant* C *independent of* u *and* τ *such that the following are valid for* $k = 1, \dots, m$:

(i) *If* $1 < p < \frac{n+2}{n}$ *then*

$$\|u_k\|_{p,\Omega\times(\tau,T)} \le C\Big[1 + \Big\|\sum_{i=1}^m u_i(\cdot,\tau)\Big\|_{1,\Omega} + \Big\|\sum_{i=1}^m u_i\Big\|_{1,\Omega\times(\tau,T)}\Big].$$

(ii) *If* $p > \frac{n+2}{n}$ *and* $r > \frac{np}{n+2}$ *then*

$$\|u_k\|_{p,\Omega\times(\tau,T)} \le C\Big[1 + \Big\|\sum_{i=1}^m u_i(\cdot,\tau)\Big\|_{r,\Omega} + \Big\|\sum_{i=1}^m u_i\Big\|_{r,\Omega\times(\tau,T)}\Big].$$

A central role in the proof of Lemma 4.1 is played by the solution of the scalar equation

$$\begin{aligned} \frac{\partial \chi}{\partial t} - d\,\Delta\chi &= \vartheta && \text{in } \Omega\times(\tau,T), \\ \chi &= 0 && \text{on } \partial\Omega\times(\tau,T), \\ \chi(\cdot,\tau) &= 0 && \text{on } \Omega, \end{aligned} \tag{4.1}$$

where $\tau < T$ and $d \in C(\mathbf{R}_+; [a,b])$ with $0 < a \le b < \infty$. We now state some more or less well-known L^q regularity results for (4.1).

LEMMA 4.2. *Let $1 < q < \infty$ and suppose that $\vartheta \in L^q(\Omega \times (\tau,T); \mathbf{R}_+)$, where $0 \le \tau < T$. Then (4.1) has a unique solution $\chi \in W_q^{2,1}(\Omega \times (\tau,T); \mathbf{R}_+)$. If $\|\vartheta\|_{q,\Omega\times(\tau,T)} = 1$ then there exists a constant $C = C(q,T)$ independent of ϑ and τ such that $\|\chi\|_{W_q^{2,1}(\Omega\times(\tau,T))} \le C$. Furthermore, C can be chosen so that*

(i) $\|\chi(\cdot\,,T)\|_{q,\Omega} \le C$;

(ii) *if $q > \frac{n+2}{2}$ then $\|\chi\|_{\infty,\Omega\times(\tau,T)} \le C$;*

(iii) *if $1 < q < \frac{n+2}{2}$ and $1 < s < \frac{nq}{n-2(q-1)}$ then*

$$\|\chi\|_{s,\Omega\times(\tau,T)} \le C \quad \text{and} \quad \|\chi(\cdot\,,T)\|_{s,\Omega} \le C;$$

(iv) $\|\chi\|_{W_q^{1,0}(\partial\Omega\times(0,T))} \le C$.

For proof of these results, we refer to §§IV.9 and II.3 of Ladyženskaja, Solonnikov, and Uralćeva [5] and §3 of Morgan [11]. We now proceed with the proof of Lemma 4.1.

Proof of Lemma 4.1. Let $p > 1$, $q = \frac{p}{p-1}$, and $\vartheta \in L^q(\Omega \times (\tau,T); \mathbf{R}_+)$ with $\|\vartheta\|_{q,\Omega\times(\tau,T)} = 1$. Take $k \in \{1,2,\dots,m\}$ and let χ be the solution of (4.1) with $d = d_k$. Now, for $t \in (\tau,T]$ define $\varphi(\cdot\,,t) = \chi(\cdot\,,T+\tau-t)$ and $\overline{\vartheta}(\cdot\,,t) = \vartheta(\cdot\,,T+\tau-t)$ so that φ satisfies

$$\begin{aligned} \frac{\partial\varphi}{\partial t} + d_k\,\Delta\varphi &= -\overline{\vartheta} && \text{in } \Omega \times (\tau,T), \\ \varphi &= 0 && \text{on } \partial\Omega \times (\tau,T), \\ \varphi(\cdot\,,T) &= 0 && \text{on } \Omega. \end{aligned}$$

We integrate $\overline{\vartheta}\sum_{i=1}^k \alpha_{k,i}u_i$ over $\Omega \times (\tau,T)$ and obtain

$$\begin{aligned} \int_\tau^T\!\!\int_\Omega \overline{\vartheta}\sum_{i=1}^k \alpha_{k,i}u_i \;\le\; & \int_\Omega \varphi(\cdot\,,\tau)\sum_{i=1}^k \alpha_{k,i}u_i(\cdot\,,\tau) + K\int_\tau^T\!\!\int_\Omega \Big(1+\sum_{i=1}^m u_i\Big)\varphi \\ & + \int_\tau^T\!\!\int_\Omega \Delta\varphi \sum_{i=1}^k \alpha_{k,i}(d_k-d_i)u_i - \int_\tau^T\!\!\int_{\partial\Omega} \frac{\partial\varphi}{\partial \mathrm{n}} \sum_{i=1}^k \alpha_{k,i}d_i g_i. \end{aligned}$$

Now, with $1 \le r \le \infty$ and $s = \frac{r}{r-1}$, Hölder's inequality gives

$$\begin{aligned} \int_\tau^T\!\!\int_\Omega \overline{\vartheta}\sum_{i=1}^k \alpha_{k,i}u_i \;\le\; C\Bigg(& \Big\|\sum_{i=1}^k u_i(\cdot\,,\tau)\Big\|_{r,\Omega} \|\varphi(\cdot\,,\tau)\|_{s,\Omega} \\ & + \Big\|1+\sum_{i=1}^m u_i\Big\|_{r,\Omega\times(\tau,T)} \|\varphi\|_{s,\Omega\times(\tau,T)} \\ & + \Big\|\sum_{i=1}^{k-1} u_i\Big\|_{p,\Omega\times(\tau,T)} \|\Delta\varphi\|_{q,\Omega\times(\tau,T)} + \Big\|\frac{\partial\varphi}{\partial \mathrm{n}}\Big\|_{q,\partial\Omega\times(\tau,T)}\Bigg) \end{aligned} \tag{4.2}$$

for some $C > 0$. If $1 < p < \frac{n+2}{n}$ then $q > \frac{n+2}{2}$, and so by Lemma 4.2 we can take $r = 1$ and $s = \infty$ and obtain by duality that

$$\|u_k\|_{p,\Omega\times(\tau,T)} \le C_p\Big(1 + \Big\|\sum_{i=1}^k u_i(\cdot\,,\tau)\Big\|_{1,\Omega} + \Big\|\sum_{i=1}^m u_i\Big\|_{1,\Omega\times(\tau,T)} + \Big\|\sum_{i=1}^{k-1} u_i\Big\|_{p,\Omega\times(\tau,T)}\Big)$$

for some $C_p > 0$. From this follows part (i) of the lemma by induction on k. Now suppose that $p > \frac{n+2}{n}$ and $r > \frac{np}{n+2}$. Then we have $q < \frac{n+2}{2}$ and $s < \frac{np}{np-(n+2)} = \frac{nq}{n-2(q-1)}$. So, from (4.2) and Lemma 4.2 we have, by duality, that

$$\|u_k\|_{p,\Omega\times(\tau,T)} \le C_p\Big(1 + \Big\|\sum_{i=1}^{k} u_i(\cdot,\tau)\Big\|_{r,\Omega} + \Big\|\sum_{i=1}^{m} u_i\Big\|_{r,\Omega\times(\tau,T)} + \Big\|\sum_{i=1}^{k-1} u_i\Big\|_{p,\Omega\times(\tau,T)}\Big)$$

for some $C_p > 0$. Part (ii) of the lemma now follows by induction on k. □

5. The proof of Theorem 1.2. We begin this section with one more lemma.

LEMMA 5.1. *Assume* (A1)–(A5). *There exist sequences* $\{C_k\}_{k=1}^{\infty} \subset (0,\infty)$ *and* $\{p_k\}_{k=1}^{\infty} \subset [1,\infty)$ *with* $p_k \uparrow \infty$ *such that, if* u *satisfies* (1.1) *with* $u_0 \in \Lambda_T$, *then for* $i = 1,\dots,m$ *and* $k \in \mathbf{N}$ *we have* $\|u_i\|_{p_k,\Omega\times(t_k,T)} \le C_k$, *where* $t_k = (1-2^{1-k})T$.

Proof. First we take $p_1 = 1$ and use the C_1 from Lemma 3.1. By that same L^1 estimate there is a $\tau_1 \in (0, \frac{1}{2}T)$ such that

$$\|u_i(\cdot,\tau_1)\|_{1,\Omega} \le \frac{C_1}{T/2}, \quad i = 1,\dots,m.$$

Now set $p_2 = \big(\frac{n+2}{n}\big)^{3/4}$. By part (i) of Lemma 4.1 there exists a C_2 such that

$$\|u_i\|_{p_2,\Omega\times(\tau_1,T)} \le C_2, \quad i = 1,\dots,m,$$

and thus $\|u_i\|_{p_2,\Omega\times(\frac{1}{2}T,T)} \le C_2$, $i = 1,\dots,m$. Therefore, there is a $\tau_2 \in (\frac{1}{2}T, \frac{3}{4}T)$ such that

$$\|u_i(\cdot,\tau_2)\|_{p_2,\Omega} \le \frac{C_2}{(T/4)^{1/p_2}}, \quad i = 1,\dots,m.$$

Now in part (ii) of Lemma 4.1 we take $r = p_2$ and $p = p_3 \equiv \big(\frac{n+2}{n}\big)^{3/2}$ and obtain a C_3 so that $\|u_i\|_{p_3,\Omega\times(\tau_2,T)} \le C_3$, $i = 1,\dots,m$ and, consequently, $\|u_i\|_{p_3,\Omega\times(\frac{3}{4}T,T)} \le C_3$, $i = 1,\dots,m$. Now we can choose $\tau_3 \in (\frac{3}{4}T, \frac{7}{8}T)$ such that

$$\|u_i(\cdot,\tau_3)\|_{p_3,\Omega} \le \frac{C_3}{(T/8)^{1/p_3}}, \quad i = 1,\dots,m,$$

and, similarly, obtain a C_4 such that $\|u_i\|_{p_4,\Omega\times(\frac{7}{8}T,T)} \le C_4$, $i = 1,\dots,m$, where $p_4 \equiv \big(\frac{n+2}{n}\big)^2$. Continuing in this way, we take $p_k = \big(\frac{n+2}{n}\big)^{k/2}$ for $k = 5,6,\dots$ and obtain a corresponding C_k such that $\|u_i\|_{p_k,\Omega\times((1-2^{1-k})T,T)} \le C_k$, $i = 1,\dots,m$. □

The preceding lemma gives rise to the following key result.

COROLLARY 5.2. *Assume* (A1)–(A5). *There exist* $C^* > 0$ *and* $t^* \in (0,T)$ *such that*

$$\|u_i\|_{\infty,\Omega\times(t^*,T)} \le C^*, \quad i = 1,\dots,m$$

for all u *satisfying* (1.1) *with* $u_0 \in \Lambda_T$.

Proof. Suppose that u satisfies (1.1) with $u_0 \in \Lambda_T$. By the polynomial growth assumption (A5) we can choose k sufficiently large in Lemma 5.1 so that each $f_i(\cdot,\cdot,u)$ is in $L^{(n+2)/2}\big(\Omega \times ((1-2^{1-k})T, T)\big)$ and, at the same time, each u_i is in $L^2\big(\Omega \times ((1-2^{1-k})T, T)\big)$ with each norm bounded independent of u. Consequently, we can apply Theorem III.8.1 of Ladyženskaja, Solonnikov, and Uralcéva [8] to obtain the desired result, where $t^* = (1-2^{-k})T$. □

We are now ready to complete the following proof.

Proof of Theorem 1.3. From Corollary 5.2 it follows that

$$\|u_i(\cdot, T)\|_{\infty,\Omega} \leq C^*, \quad i = 1, \ldots, m$$

for all u satisfying (1.1) with $u_0 \in \Lambda_T$. Thus, by (2.2) there is a constant $\widetilde{C}$ such that

$$\|u_{0_i}\|_{\infty,\Omega} \leq \widetilde{C}, \quad i = 1, \ldots, m$$

for all $u_0 \in \Lambda_T$. That is, Λ_T is a bounded subset of $C(\overline{\Omega}; \mathbb{R}^m_+)$. So, by Theorem 2.1 and the discussion in §2, the mapping F defined by (2.1) has a fixed point $z^* \in C_0(\overline{\Omega}; \mathbb{R}^m)$, which gives rise to a fixed point $u_0^* = z^* + \tilde{g} \in C(\overline{\Omega}; \mathbb{R}^m_+)$ of $\mathcal{S}(T)$. Now, by the periodicity of f and g and the uniqueness of solutions to (1.1), it follows that (1.1) possesses the T-periodic solution $u(\cdot, t) = \mathcal{S}(t)u_0^*$. □

6. Remarks and generalizations. Straightforward modifications of our proofs show that Theorem 1.2 remains valid if the boundary conditions are of Robin type with smooth, T-periodic parameters. If the boundary conditions are of Neumann type, then (A5) must be modified so that

$$\sum_{j=1}^{m} \alpha_{m,j} f_j(\cdot, \cdot, u) \leq K - \epsilon \sum_{j=1}^{m} u_j \tag{6.1}$$

for some $\epsilon > 0$ in order to obtain the L^1 estimate in Lemma 3.1. Also, if boundary conditions are of Dirichlet or Robin type, then (A5) can be replaced by

$$\sum_{j=1}^{m} \alpha_{m,j} f_j(\cdot, \cdot, u) \leq K + \epsilon \sum_{j=1}^{m} u_j,$$

where $\epsilon > 0$ provided that ϵ is sufficiently small. (However, it is generally crucial that the boundary condition *type* be uniform throughout the system.) For both the Neumann and Robin cases, the operator F in §2 would be given simply by $Fz = \mathcal{S}(T)z^+$, mapping $C(\overline{\Omega}; \mathbb{R}^m)$ into itself, and the set Λ_T would consist of all u_0 satisfying $u_0 = \sigma \mathcal{S}(T)u_0$ with $0 < \sigma < 1$.

Certainly one would like to allow x dependence in the diffusivities; i.e., have operators of the form $\nabla \cdot (d_i(x,t)\nabla u_i)$ in (1.1). Assuming smoothness and uniform ellipticity, the only obstacle to this goal is the L^1 estimate in Lemma 3.1. If such an estimate were available then the remainder of the argument would proceed with only minor modification. This estimate is readily available in the case of Neumann or Robin boundary conditions provided that f satisfies (6.1). One arrives at this estimate by setting $w = \sum_{k=1}^{m} \alpha_{m,k} u_k$ and integrating

$$\frac{\partial w}{\partial t} \leq \nabla \cdot \sum_{k=1}^{m} \alpha_{m,k} d_k \nabla u_k + K - \tilde{\epsilon} w$$

over $\Omega \times (0, T)$ with $u_0 = \sigma u(\cdot, T)$ and $0 < \sigma < 1$. Here $\tilde{\epsilon} = \epsilon \min\{\alpha_{m,1}, \ldots, \alpha_{m,m}\}$.

We also remark that if all the diffusivities are equal, i.e., $d_1 = d_2 = \cdots = d_m$, then we need neither the intermediate sums condition (A1) nor the polynomial growth

condition (A2) to prove Theorems 1.1 and 1.2. Indeed, in this case, global existence follows easily from (A5). By introducing, if necessary, the additional equation

$$\frac{\partial u_{m+1}}{\partial t} - d_1 \Delta u_{m+1} = K - \sum_{j=1}^{m} \alpha_{m,j} f_j(\cdot, \cdot, u)$$

into (1.1) along with zero boundary and initial values, we can assume without loss of generality that $\sum_{j=1}^{m} \alpha_{m,j} f_j(\cdot, \cdot, u) = K$, and so $w \equiv \sum_{k=1}^{m} \alpha_{m,k} u_k$ satisfies

$$\begin{aligned} \frac{\partial w}{\partial t} - d_1 \Delta w &= K && \text{in } \Omega \times (0,T), \\ w &= \sum_{k=1}^{m} \alpha_{m,k} g_k && \text{on } \partial\Omega \times (0,T), \\ w(\cdot,0) &= \sum_{k=1}^{m} \alpha_{m,k} u_{0_k} && \text{on } \overline{\Omega}. \end{aligned}$$

By applying Lemma 4.1 to this scalar equation and using Lemma 3.1 and the argument in the proof of Lemma 5.1, one finds a $t^* \in (0,T)$ such that $\|w\|_{2,\Omega\times(t^*,T)} \leq C$, where C is independent of $u_0 \in \Lambda_T$. From this fact and Theorem III.8.1 of Ladyženskaja, Solonnikov, and Uralcéva [8], we arrive at the necessary estimate on $\|w(\cdot,T)\|_{\infty,\Omega}$.

REFERENCES

[1] H. AMANN, *Periodic solutions of semilinear parabolic equations*, in Nonlinear Analysis, A Collection of Papers in Honor of E. H. Rothe, Academic Press, New York, 1978.

[2] D. BANGE, *An existence theorem for periodic solutions of a nonlinear parabolic boundary value problem*, J. Differential Equations, 24 (1977), pp. 426–436.

[3] R. GAINES AND W. WALTER, *Periodic solutions to nonlinear parabolic differential equations*, Rocky Mountain J. Math., 7 (1977), pp. 297–312.

[4] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1970.

[5] S. HOLLIS, R. MARTIN, AND M. PIERRE, *Global existence and boundedness in reaction-diffusion systems*, SIAM J. Math. Anal., 18 (1987), pp. 744–761.

[6] S. HOLLIS AND J. MORGAN, *Partly dissipative reaction-diffusion systems and a model of phosphorus diffusion in silicon*, Nonlinear Anal., 19 (1992), pp. 427–440.

[7] ———, *On the blow-up of solutions to some semilinear and quasilinear reaction-diffusion systems*, Rocky Mountain J. Math., 24 (1994), pp. 1447–1465.

[8] O. LADYŽENSKAJA, V. SOLONNIKOV, AND N. URALĆEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[9] B.-P. LIU AND C. V. PAO, *Periodic solutions of coupled semilinear parabolic boundary value problems*, Nonlinear Anal., 6 (1982), pp. 237–252.

[10] J. MORGAN, *Global existence for semilinear parabolic systems*, SIAM J. Math. Anal., 20 (1989), pp. 1128–1144.

[11] ———, *Boundedness and decay results for reaction-diffusion systems*, SIAM J. Math. Anal., 21 (1990), pp. 1172–1189.

[12] J. MURRAY, *Mathematical Biology*, Springer-Verlag, New York, 1989.

[13] F. ROTHE, *Global Solutions of Reaction-Diffusion Systems*, Lecture Notes in Math. 1072, Springer-Verlag, Berlin, 1980.

[14] T. SEIDMAN, *Periodic solutions of a nonlinear parabolic equation*, J. Differential Equations, 19 (1975), pp. 245–257.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1233–1256, September 1995

009

BOUNDEDNESS OF SOLUTIONS FOR QUASIPERIODIC POTENTIALS*

M. LEVI† AND E. ZEHNDER‡

Abstract. In this paper conservative systems are studied describing the motion of a particle on the line in the field of a potential force with additional quasiperiodic time dependence.

It is shown that superquadratic growth of the potential at infinity results in the near-integrability of the Hamiltonian system in question (for a large class of potentials), despite the fact that no smallness assumptions are made on the quasiperiodic dependence of the potential on time. As a consequence all the solutions of such systems are bounded for all time. Some specific examples are given, together with a counterexample which shows that, without the quasiperiodicity assumption, the boundedness breaks down.

Key words. quasiperiodicity, stability, action-angle variables, normal term, KAM theory

AMS subject classifications. 34, 58

1. Introduction and results.

We shall study the boundedness of all solutions of time-dependent equations having the form

$$\ddot{x} + V_x(x, \omega t) = 0, \quad x \in \mathbf{R}. \tag{1}$$

This equation is the simplest yet highly nontrivial model of conservative systems such as charged particles in periodic fields. Setting $\dot{x} = y$, these equations can be rewritten in Hamiltonian form with Hamiltonian functions

$$H(x, y, t) = \frac{1}{2}y^2 + V(x, \omega t) \tag{2}$$

on the extended phase space $(x, y, t) \in \mathbf{R}^3$.

A distinguished class of equations describe forced pendulum-like systems in which the potential is assumed to be periodic in x such that V is a function on $S^1 \times \mathbf{R}$. If, in addition, the time dependence is periodic or even quasiperiodic, then it turns out that every solution $x(t)$ of (1) is bounded in the phase space $S^1 \times \mathbf{R}$, i.e., $\sup\{|\dot{x}(t)|, t \in \mathbf{R}\} < \infty$, provided that only the potential V is sufficiently smooth and, in the quasiperiodic case, the frequencies meet a Diophantine condition. The proof of this phenomenon is based on the observation that such systems are near so-called integrable systems in the region of the phase space $S^1 \times \mathbf{R}$ in which y is sufficiently large. For proofs we refer to Levi [12], Moser [13], and Chierchia and Zehnder [14]. If, however, the smoothness requirements are not met, for example, if V merely belongs to the class C^2, then unbounded solutions may be expected. Also, if the above restrictions on the time dependence are dropped, unbounded solutions are likely to occur even for smooth and bounded potentials V.

If the periodicity requirement in x is dropped, then the configuration space is no longer S^1 but $\mathbf{R}^1$, and the question of boundedness of all solutions for (1) is much more subtle. It is related to the asymptotic behavior of the nonlinearity in x, the

* Received by the editors May 14, 1993; accepted for publication (in revised form) February 11, 1994.

† Department of Mathematics, Rensselaer Polytechnic Institute, Troy, New York 12180.

‡ Forschungsinstitut für Mathematik, Eidgenössische Technische Hochschule-Zentrum, 8092 Zürich, Switzerland.

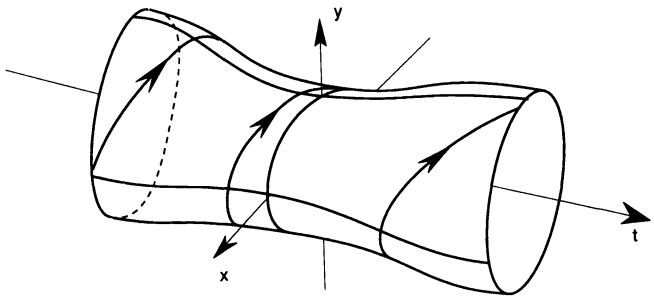


FIG. 1.1.

smoothness of V, and also the nature of the time dependence. For example, if

$$H(x,y,t) = \frac{1}{2}y^2 + \frac{1}{\sqrt{x^2 + r(t)}}$$

for a time periodic and positive function $r(t)$, as in the case of the so-called restricted three-body problem, then the solutions with initial conditions $\dot{x}(0) = y(0)$ sufficiently large are clearly not bounded. Here the level lines of $H(x,y,t) = E$ for frozen t are not closed curves if E is large. In contrast, in the example

$$\ddot{x} + a(t)x^3 + b(t)x^2 + c(t)x = p(t) \tag{3}$$

level lines for E large are closed curves; see Fig. 1.1. However, the energy is not conserved in time and might increase, forcing a solution to be unbounded in the phase space $\mathbf{R}^2$. For this class of examples the subtle question of boundedness of solutions was already raised by Littlewood, who constructed examples [2] admitting unbounded solutions assuming a periodic but discontinuous forcing $p(t)$; see also Levi [3] and Long [4]. Recently Zharnitsky [30] succeeded in constructing such an example with $p(t)$ discontinuous. In 1976 Morris [5] succeeded in proving that, for a continuous time periodic forcing, all the solutions of

$$\ddot{x} + x^3 = p(t)$$

are bounded in $\mathbf{R}^2$. For more recent results in the time periodic case we refer to [6]–[9], [16]. We also mention the related problem of Ulam and Fermi's "ping-pong," consisting of a particle bouncing between a wall and a periodically moving "paddle" parallel to the wall, undergoing perfectly elastic collisions with both; a basic physical question is whether the energy of a particle stays bounded for all time in such a periodically varying system. It should be mentioned that this problem is a limiting case of the aforementioned problems where the walls of the potential well become infinitely steep. The affirmative answer to the last question for sufficiently smooth periodic motions of the wall has been given by Moser (in an unpublished, private communication), Douady [27], and in [6]. In what follows we shall assume that the time dependence in (1) is quasiperiodic with frequencies $\omega = (\omega_1, \ldots, \omega_N) \in \mathbf{R}^N$; i.e., we assume

$$V(x,\omega t) = V(x,\omega_1 t, \ldots, \omega_N t), \tag{4}$$

where $V(x,\xi_1,\ldots,\xi_N)$ is assumed to be periodic of period 1 in all the variables $\xi_1,\ldots,\xi_N$. Moreover, we shall assume that the frequencies ω are not only rationally

independent, but meet the Diophantine conditions

$$|\langle \omega, j \rangle| \geq \gamma |j|^{-\tau} \quad \text{for all } j \in \mathbf{Z}^N \backslash \{0\} \tag{5}$$

with two constants $\tau > N$ and $\gamma > 0$. The brackets on the left-hand side denote the scalar product.

The system considered first is of the form

$$V(x, \omega t) = \sum_{j=1}^{2n+2} a_j(\omega t) x^j, \tag{6}$$

$n \geq 1$, where all the coefficients are quasiperiodic functions in time with the same frequencies $\omega \in \mathbf{R}^N$ and, in addition, the leading coefficient $a = a_{2n}$ is positive:

$$a(\omega t) \geq \min_{\xi \in T^N} a(\xi) > 0. \tag{7}$$

1.1. Stability and invariant tori for polynomial potentials.

THEOREM 1. *Let the polynomial potential $V(x, \omega t)$ satisfy* (6) *and* (7) *together with the Diophantine conditions* (5), *and assume that $a_j \in C^k(T^N)$ for $k > 4\tau + 6$ and $0 \leq j \leq 2n + 2$. Then all solutions $x(t)$ of $\ddot{x} + V_x(x, \omega t) = 0$ are bounded, i.e.,*

$$\sup_{t \in \mathbf{R}} (x(t)^2 + \dot{x}(t)^2) < \infty.$$

Note that the smoothness requirement depends only on the number of the underlying frequencies ω and not on the degree of the polynomial in x. Already for the time periodic case this statement is not quite obvious. It has only recently been proved by Laederich and Levi in [6], improving and simplifying an earlier result of Dieckerhoff and Zehnder [7]. We should point out that all the boundedness proofs in the time periodic case use Moser's twist theorem [19] and its regularity improvements [10] and [11] in a crucial way.

In order to describe the idea of the proof in the quasiperiodic case, we write the equation as a system in the phase space $\mathbf{R}^3$,

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -V_x(x, t), \\ \dot{t} &= 1, \end{aligned} \tag{8}$$

and abbreviate the vector field in $\mathbf{R}^3$ on the right-hand side by X. For $C > 0$ we denote by A_C the region $A_C = \{(x, y, t) | x^2 + y^2 > C\}$ in $\mathbf{R}^3$. For every $C > 0$ we shall construct an embedded cylinder $w : \mathbf{R} \times S^1 \to \mathbf{R}^3$ contained in A_C,

$$w : (t, s) \mapsto (u(t, s), v(t, s), t), \tag{9}$$

satisfying

$$C < \inf_{\mathbf{R} \times S^1} (u^2 + v^2) < \sup_{\mathbf{R} \times S^1} (u^2 + v^2) < \infty$$

and which is tangential to the vector field X in $\mathbf{R}^3$, so that it is invariant under the flow of X. Now if a solution $(x(t), y(t), t)$ of (8) satisfies $x(t^*)^2 + y(t^*)^2 \leq C$ for some

$t^* \in \mathbf{R}$, then it follows from the invariance of the cylinder and the uniqueness of the solutions that this solution does exist for all times $t \in \mathbf{R}$ and satisfies, in addition, $(x(t)^2 + y(t)^2) \leq \sup_{\mathbf{R}\times S}(u^2 + v^2) < \infty$ for all $t \in \mathbf{R}$. Hence it is bounded.

The existence of these invariant surfaces in $\mathbf{R}^3$ will be concluded from the observation that the system (8) is in the region A_C near an integrable system, provided that C is sufficiently large, so that well-known small denominator perturbation techniques can be applied. These techniques require an excessive amount of smoothness; this is, of course, well known. The near integrability is, however, not obvious a priori and its proof is the main task. It will be apparent only after scaling the time t and the phase space variables and only after several coordinate changes, which transform the vector field into a suitable form.

The invariant surface found consists of quasiperiodic solutions, and we shall prove the following existence statement.

THEOREM 2. *The equation*

$$\ddot{x} + V_x(x, \omega t) = 0$$

with the potential V satisfying the assumptions of Theorem 1 possesses uncountably many quasiperiodic solutions having $1 + N$ frequencies $(\alpha, \omega) \in \mathbf{R}^{1+N}$ and satisfying the Diophantine conditions

$$|\alpha k + \langle \omega, j \rangle| \geq \gamma(|k| + |j|)^{-\tau}$$

for all $(k, j) \in \mathbf{Z}^{1+N}\backslash\{0\}$ with the same constants $\gamma > 0$ and $\tau > N$ as in (5).

Indeed, as expected, the dominant part of the phase space for $x^2 + \dot{x}^2$ large is covered by quasiperiodic solutions. The worst possible failure of the Diophantine condition (5) corresponds to all frequencies ω being rational multiples of one of them; in this case V is time periodic and the aforementioned results for the periodic case apply, showing boundedness under appropriate assumptions.

The intermediate (Liouville) case between the periodic one and the Diophantine one is less clear. It seems highly likely that when ω is a Liouville vector, where (5) fails for infinitely many j, an arbitrarily small change in V would destroy an invariant surface with fixed frequencies if there is one; see Mather [21]. On the other hand, it is less clear whether all such tori can be destroyed at once, or perhaps the destruction of one torus could lead to the creation of another one.

1.2. General potentials. So far we have considered a rather restricted class of potentials. It turns out that the ideas of the proof of the aforementioned Theorems can be applied to a more general class of quasiperiodic potentials introduced recently in [15] for the time periodic case.

THEOREM 3. *If ω satisfies the Diophantine conditions (5) then there exist positive constants $a, b, \mu_0(b)$, and $\gamma(b)$ such that the conclusions of Theorems 1 and 2 hold for the equation*

$$\ddot{x} + V_x(x, \omega t) = 0$$

provided $V(x, \xi)$ with $x \in \mathbf{R}$ and $\xi \in T^N$ belongs to $C^d, d = 4\tau + 7 + \gamma(b)$ and, moreover, satisfies the following conditions:

(i)

$$V(x, \xi) \to \infty \text{ as } |x| \to \infty$$

uniformly in $\xi \in T^N$.

(ii) *In the notation*

$$W := \frac{V}{V_x} \text{ and } U := \frac{V_\xi}{V_x},$$

the following estimates hold for all $(x,\xi) \in \mathbf{R} \times \mathbf{R}^N$ *and* $|x|$ *large:*

(iia) $$-\frac{1}{2} + a < \partial_x W(x,\xi) < \frac{1}{2} - b, \quad b < 1 - a,$$

(iib) $$|\partial_x^k \partial_\xi^\alpha V(x,\xi)| \le C_{k\alpha} |x|^{-k} |V(x,\xi)|^{1+\mu} \quad \text{for some} \quad 0 \le \mu \le \mu_0(b)$$

and

$$|\partial_x^k \partial_\xi^\alpha U|, |\partial_x^k \partial_\xi^\alpha W| \le C_{k\alpha} |x|^{1-k}$$

for all $|k| + |\alpha| \le 4\tau + 6 + \gamma(b)$.

This theorem will be proved in §4.

Examples. The above conditions (i) and (ii) with $\mu = 0$ hold for polynomial (in x) potentials. With $\mu \neq 0$ the class of potentials widens considerably to include exponential growth, oscillatory growth, and much more. The simplest example is $V_1(x,t) = p(\omega t)\cosh x$; it satisfies all the conditions if $p > 0$ is smooth enough.

A more difficult potential

$$V_2(x,t) = p(\omega t)(\cosh x + q(x)),$$

where q is any polynomial, satisfies conditions (i) and (ii) as well, again provided $p : T^n \to \mathbb{R}$ is smooth enough. The polynomial $q(x)$ in the above example can be replaced by, say, $\cos x$ or a polynomial in x and $\cos x$ without violating the conditions of Theorem 3:

$$V_3(x,t) = p(\omega t)(\cosh x + q_1(x, \cos x)).$$

Yet another example is

$$V_4(x,t) = \cosh\left[(3 + \cos t + \cos\sqrt{2}t)(x + \sin(1+x^2)^\nu)\right]$$

with $\nu > 0$ sufficiently (specifically) small.

The list can be continued indefinitely.

It should be emphasized that analogous stability results cannot be expected in higher dimensions. Consider, for example, a time-dependent Hamiltonian system defined by the Hamiltonian function

$$H(x,y,\omega,t) = \frac{1}{2}|y|^2 + V(x,\omega t)$$

on $(x,y) \in T^n \times \mathbb{R}^n$ for which the energy is not conserved. Here one also finds an abundance of quasiperiodic solutions in the region of the phase space where $|y|$ is large, in which the system can be considered as a system near an integrable one, provided V is sufficiently smooth; see, e.g., [14]. However, if $n > 1$ then the existence of these solutions does not lead to bounds for all solutions of the system. But there are bounds

for all solutions not over an infinite interval of time but over an exponentially large interval of time, provided the potential is not only smooth but real analytic. This well-known phenomenon has been discovered by Nekhoroshev [25]. As an illustration we mention the effective bounds for the above example. Assume $V(t,x)$ with $(t,x) \in \mathbb{R}^n \times T^n$ has a holomorphic extension to an imaginary strip $|\mathrm{Im}\, t| \leq \sigma$ and $|\mathrm{Im}\, t| \leq \sigma$ for some positive number σ. Then there are positive constants T^* and R^* depending on V, σ, and the dimension n such that, for every $\rho \geq R^*$ and every solution $(x(t), y(t))$ of the Hamiltonian equation, we have

$$|y(t) - y(0)| \leq \rho$$

for all t in the interval

$$|t| \leq T^* \exp\left(\frac{\rho}{R^*}\right)^{\alpha}.$$

The proof of these estimates with explicit constants is based on Nekhoroshev's ideas, and we refer to [26] (with $\alpha = 2/(n^2+n)$) and [28] and [29] for the recently improved estimate with $\alpha = \frac{1}{2n}$. This is merely a special example of an exponential stability result which replaces the stronger stability results of Theorem 3 for systems in higher dimensions.

1.3. Unbounded solutions with nonrecurrent forcing. We shall show that, as soon as the quasiperiodicity requirement of the time dependence is removed, even the "nicest" equations can have unbounded solutions for forcing terms which are smooth, small, and tend to zero as the time goes to infinity.

THEOREM 4. *Given any $\varepsilon > 0$ and any $r \in \mathbf{N}$, there exists a function $p \in C^\infty(\mathbf{R})$ satisfying*

$$\|p\|_{C^r(\mathbf{R})} \leq \varepsilon \text{ and } \lim_{t\to\infty} D^j p(t) = 0 \text{ for } 0 \leq j \leq r-1, \tag{10}$$

such that the equation

$$\ddot{x} + x^3 = p(t) \tag{11}$$

possesses an unbounded solution $y(t)$. Moreover, the rate of decay in (10) *is given by*

$$\sup_{t>0} t^{\frac{2(r-j)}{2r+3}} |D^j p(t)| < \infty \tag{12}$$

for $0 \leq j \leq r-1$, and the rate of growth of the unbounded solution $y(t)$ is given by

$$\frac{1}{C} t^{\frac{4}{2r+3}} \leq \frac{1}{2}\dot{y}(t)^2 + \frac{1}{4} y(t)^4 \leq C t^{\frac{4}{2r+3}} \tag{13}$$

for $t \geq 1$ with a positive constant C depending on ε.

It should be pointed out that the first part of the theorem holds true for every equation $\ddot{x} + V_x(x) = p(t)$ provided that $\frac{1}{x} V_x(x) \to \infty$ as $|x| \to \infty$; i.e., an unbounded solution can be produced with a forcing satisfying (10). This will follow from the proof. As for another, more subtle phenomenon we recall that Coffman and Ullrich [22] constructed a positive and continuous function $p(t)$ close to a constant, which, however, is not of bounded variation near a point t^*, such that the equation $\ddot{x} + p(t)x^3 = 0$ has a solution which is unbounded on the finite interval $0 \leq t < t^*$.

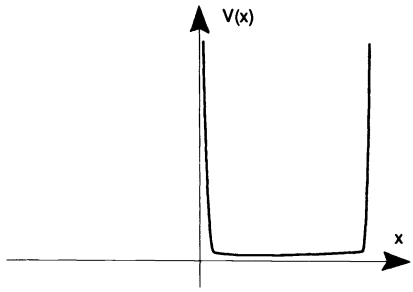


FIG. 2.1.

2. A "squash player's" potential and some open problems. Let $M > 0$ be a large integer and consider equation (1) with the special potential

$$V(x, \omega t) = (x-1)^{2M} + \sum_{i=1}^{n} (x/p_i(\omega_i t))^{2M},$$

where $0 < p_i(\tau) < 1$ are periodic functions of period 1. Since M is large, V has two steep "walls," one near $x = 0$ and the other near $x = \min_{1 \le i \le n} \{p_j(\omega_i t)\}$; see Fig. 2.1. One may think of n squash players each moving his racket periodically, holding it at the distance $p_i(\omega_i t)$ from the wall $x = 0$. The player whose racket is in front, i.e., closest to the wall at a given moment, "gets to hit the ball." It would make sense to assume that $\min_\tau p_i(\tau) < \max_\tau p_j(\tau)$ for all $i, j = 1, \ldots, n$, so that everyone gets a chance to hit. Now Theorem 1 shows that as long as the frequency vector ω is Diophantine and $p_i \in C^k(S^1)$ with $k > 4\tau + 6$, the game will proceed without an escalation, i.e., both the speed and the position of any possible motion will stay bounded for all $t \in \mathbf{R}$. Of course, no explicit estimate on that bound is given and no estimate is given on how deep the potential wall is penetrated.

Open problems. 1. Modifying the "squash" example by making the walls rigid, we obtain a problem not covered by Theorems 1 or 2. In fact, it is doubtful that the result still holds for such a modification since the smoothness of the "potential" is lost in taking the rigid limit.

2. As for a different modification, one could consider the Ulam–Fermi "ping-pong" problem consisting of a particle bouncing elastically between two parallel walls with the walls undergoing a quasiperiodic motion. The problem is to prove that the velocity of every motion is bounded for all time provided the motion of the walls is smooth enough and the Diophantine conditions hold by establishing the existence of invariant cylinders in the extended phase space of the system.

3. Proof of Theorems 1 and 2. We first transform the equation

$$\ddot{x} + V_x(x, \omega t) = 0, \quad x \in \mathbf{R} \tag{14}$$

with V satisfying assumptions (4)–(7) into a suitable form. We proceed in several steps.

3.1. The rescaling into a slow system. As in [6] we first rescale the time variable t and, at the same time, the space variable x setting for small $\delta > 0$,

$$u = \delta x, \; t = \varepsilon s, \quad \text{where } \varepsilon = \delta^n. \tag{15}$$

If $x(t)$ is a solution of (14), then

$$u(s) := \delta x(\varepsilon s)$$

is a solution of the equation

$$\frac{d^2}{ds^2}u + \varepsilon^2 \delta V_x\left(\frac{u}{\delta}, \varepsilon\omega s\right) = 0, \quad u \in \mathbf{R}. \tag{16}$$

In view of the assumptions on V we are led to the equivalent differential equation

$$\frac{d^2}{ds^2}u + W_u(u, \varepsilon\omega s, \varepsilon) = 0, \tag{17}$$

where $W(u, \xi, \varepsilon)$, $\xi \in T^N$ is given by

$$W(u, \xi, \varepsilon) = a(\xi)u^{2n+2} + \varepsilon^\alpha \sum_{j=1}^{2n+1} a_j(\xi)\varepsilon^{\alpha(2n+1-j)}u^j \quad \text{with } \alpha = \frac{1}{n}$$

and $a \equiv a_{2n+2}$; moreover, $a_j \in C^k(T^N)$. Now, returning to the old notation by replacing u by x and s by t we therefore arrive at the Hamiltonian system

$$H(x, y, \varepsilon\omega t, \varepsilon) = \frac{1}{2}y^2 + W(x, \varepsilon\omega t, \varepsilon) \tag{18}$$

in the extended phase space $(x, y, t) \in \mathbf{R}^3$. Our aim is to construct, for every $\varepsilon > 0$, an invariant cylinder for (18) contained in $(x, y, t) \in A \times \mathbf{R}$ having the time axis in its interior, where A is a fixed and bounded annular region in $\mathbf{R}^2$ around the origin.

3.2. The action-angle variables. At first we consider the time-independent Hamiltonian system in $(x, y) \in \mathbf{R}^2$ given by

$$H(x, y, \xi, \varepsilon) = \frac{1}{2}y^2 + W(x, \xi, \varepsilon), \tag{19}$$

which depends on the parameters $\xi \in \mathbf{R}^N$ and $\varepsilon > 0$. The dependence on each ξ_i is periodic with each period 1. If ε is sufficiently small one can introduce in an annulus-like domain in the (x, y)-plane so-called action and angle variables $(\varphi, I) \mapsto (x, y)$, using a generating function $S(x, I) \equiv S(x, I, \xi, \varepsilon)$ depending periodically on ξ by the formula

$$\begin{aligned} y &= S_x(x, I), \\ \varphi &= S_I(x, I). \end{aligned} \tag{20}$$

As usual [23], the action variable I is defined as the area of the level curve γ in $\mathbf{R}^2$ defined by $H(x, y, \xi, \varepsilon) = E$:

$$I = \int_\gamma y dx = I(E, \xi, \varepsilon). \tag{21}$$

If E exceeds all critical values of W in x, then γ is a simple closed curve. Since $\frac{\partial I}{\partial E} > 0$ for E large enough and $\varepsilon > 0$ small enough, we can define the inverse function of $I(E)$ so that E is a well-defined function of (I, ξ, ε), which we denote by K^0. The

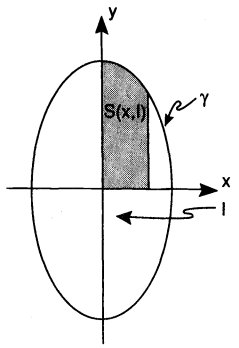


FIG. 3.1.

generating function S is then defined as the shaded area in Fig. 3.1 or, equivalently, as the solution of

$$K^0(I,\xi,\varepsilon) = H(x, S_x(x,I,\xi,\varepsilon),\xi,\varepsilon); \tag{22}$$

it is independent of the angle variable φ. As in [6] one verifies that

$$K^0(I,\xi,\varepsilon) = a(\xi)I^\beta + O(\varepsilon^\alpha) \tag{23}$$

with $\beta = \frac{2n+2}{n+2}$ and a positive function $a \in C^k(T^N)$. Now setting $\xi = \varepsilon\omega t$, we define the *time-dependent* symplectic transformation (20) by means of the generating function $S = S(x,I,\varepsilon\omega t,\varepsilon)$. It transforms the Hamiltonian system (18) into the system

$$K(\varphi,I,\varepsilon\omega t,\varepsilon) = K^0(I,\varepsilon\omega t,\varepsilon) + \frac{\partial}{\partial t}S(x,I,\varepsilon\omega t,\varepsilon) \tag{24}$$

on the phase space $(\varphi,I,t) \in S^1 \times \mathbf{R} \times \mathbf{R}$. Here $I \in \mathbf{R}$ varies in a bounded interval which is independent of ε; ε is small and $x = x(\varphi,I,\varepsilon\omega t,\varepsilon)$ in view of (20).

We denote the second term on the right-hand side by $K^1(\varphi,I,\varepsilon\omega t,\varepsilon)$, so that (24) becomes

$$K(\varphi,I,\varepsilon\omega t,\varepsilon) = K^0(I,\varepsilon\omega t,\varepsilon) + K^1(\varphi,I,\varepsilon\omega t,\varepsilon).$$

In what follows it will be crucial that

$$K^1(\varphi,I,\varepsilon\omega t,\varepsilon) = \varepsilon\omega \cdot \frac{\partial}{\partial\xi}S(x,I,\varepsilon\omega t,\varepsilon) = O(\varepsilon) \tag{25}$$

with all its (finitely many) derivatives in (φ,I,t). Here $\frac{\partial}{\partial\xi}S$ is the gradient of S with respect to ξ.

3.3. Choosing the symplectic angle as time. The Hamiltonian equations associated with the function K in (24) are, on the extended phase space $(\varphi,I,t) \in S^1 \times \mathbb{R} \times \mathbb{R}$, given by

$$\begin{aligned} \frac{d\varphi}{dt} &= K_I(\varphi,I,\varepsilon\omega t,\varepsilon), \\ \frac{dI}{dt} &= -K_\varphi(\varphi,I,\varepsilon\omega t,\varepsilon), \\ \frac{dt}{dt} &= 1. \end{aligned} \tag{26}$$

It is well known that if the time t and the "energy" K are chosen as the new conjugate variables and the angle φ is chosen as the new time variable, then (26) is transformed into an equation which is again Hamiltonian and belongs to a Hamiltonian function Q, which is the inverse of the function $I \mapsto K(\varphi, I, \varepsilon\omega t, \varepsilon)$. Indeed, from (23) we conclude that the partial derivative $K_I > 0$ if ε is small, and we can therefore define the transformation

$$\psi: \quad \begin{cases} q = t, \\ p = K(\varphi, I, \varepsilon\omega t, \varepsilon), \\ s = \varphi \end{cases} \tag{27}$$

from $(\varphi, I, t) \in S^1 \times \mathbb{R} \times \mathbb{R}$ into $(q, p, s) \in \mathbb{R} \times \mathbb{R} \times S^1$. Denote by $Q(\xi, p, \varphi, \varepsilon)$ the inverse function of $I \mapsto K(\varphi, I, \xi, \varepsilon)$, so that

$$p = K(\varphi, Q(\varepsilon\omega t, p, \varphi, \varepsilon), \varepsilon\omega t, \varepsilon). \tag{28}$$

Then the flow induced in the (q, p, s)-space by (26) is Hamiltonian with Q as the Hamiltonian function. This follows from $Id\varphi - Kdt = -(pdq - Qds)$ (see [23], [24]) or by a direct calculation which we now carry out. Abbreviating the vector field on the right-hand side of (26) by X, one readily verifies, using (27) and (28), that the transformed vector field is given by

$$\frac{d}{dt}\begin{pmatrix} q \\ p \\ s \end{pmatrix} = (d\psi)^{-1} X \circ \psi = \begin{pmatrix} 1 \\ -Q_p^{-1} Q_q \\ Q_p^{-1} \end{pmatrix},$$

where $Q = Q(\varepsilon\omega q, p, s, \varepsilon)$. Multiplying this vector field by the positive function Q_p we find as claimed

$$\begin{aligned} \frac{dq}{ds} &= Q_p(\varepsilon\omega q, p, s, \varepsilon), \\ \frac{dp}{ds} &= -Q_q(\varepsilon\omega q, p, s, \varepsilon), \end{aligned} \tag{29}$$

where the new time variable $s = \varphi$ is the old angle. One verifies, moreover, that

$$Q(\varepsilon\omega q, p, s, \varepsilon) = Q^0(\varepsilon\omega q, p, \varepsilon) + \varepsilon Q^1(\varepsilon\omega q, p, s, \varepsilon), \tag{30}$$

where

$$Q^0(\varepsilon\omega q, p, \varepsilon) = b(\varepsilon\omega q)p^\gamma + O(\varepsilon^\alpha) \tag{31}$$

with $\gamma = \frac{n+2}{2n+2}$ and a positive periodic function $b \in C^k(T^N)$.

3.4. Removing the time dependence in the dominant term. In order to relate the notation of the variables q, p, s of the Hamiltonian function (30) to their original meaning, we return to the old notation and replace the variables (q, p, s) and the Hamiltonian function Q by (t, K, ε) and I, so that the Hamiltonian (30) in this notation is

$$I(\varepsilon\omega t, K, \varphi, \varepsilon) = I^0(\varepsilon\omega t, K, \varepsilon) + \varepsilon I^1(\varepsilon\omega t, K, \varphi, \varepsilon), \tag{32}$$

where $I^0 = Q^0$ and $I^1 = Q^1$ are given by (30) and (31). The Hamiltonian equations now look as follows:

$$\frac{dt}{d\varphi} = I_K \ (\varepsilon\omega t, K, \varphi, \varepsilon), \tag{33}$$

$$\frac{dK}{d\varphi} = -I_t \ (\varepsilon\omega t, K, \varphi, \varepsilon).$$

Introducing $\varepsilon t = T$, these equations become

$$\frac{dT}{d\varphi} = \varepsilon \, I_K \ (\omega T, K, \varphi, \varepsilon), \tag{34}$$

$$\frac{dK}{d\varphi} = -\varepsilon I_T \ (\omega T, K, \varphi, \varepsilon)$$

and belong to the Hamiltonian function

$$\varepsilon \, I \, (\omega T, K, \varphi, \varepsilon).$$

We look for a symplectic transformation ψ of the form

$$\psi : \begin{array}{rcl} \tau & = & T + u(T, K), \\ h & = & K + v(T, K), \end{array} \tag{35}$$

which is, in particular, independent of the (time) variable φ and transforms the Hamiltonian function εI into the following form:

$$\varepsilon \, \{I \circ \psi^{-1}\} = \varepsilon \, \{\Phi^0(h, \varepsilon) + \varepsilon \, \Phi^1(\omega\tau, h, \varphi, \varepsilon)\}, \tag{36}$$

where the dominant term Φ^0 is independent of τ. To define this transformation we first define the leading term Φ^0 by taking the inverse function of $K \to I^0(\omega T, K)$, then averaging it over the torus T^N and taking the inverse again. Observe that $I^0_K > 0$ in the view of (31) provided that ε is sufficiently small. Therefore we can solve $I^0(\xi, K) = h$ for K and find a function $K^0(\xi, h)$ satisfying

$$I^0(\xi, K^0(\xi, h)) = h, \tag{37}$$

where K^0 is periodic in $\xi \, \epsilon \, T^N$. Next, define the mean value over the torus by

$$[K^0](h) = \int_{T^N} K^0(\xi, h) d\xi. \tag{38}$$

Since $K^0_h = (I^0_K)^{-1} > 0$, the function $[K^0]$ has an inverse Φ^0 which thus satisfies

$$[K^0] \circ \Phi^0(h) = h. \tag{39}$$

In our notation we have neglected the dependence on ε. Clearly, $\Phi^0_h > 0$. This finishes the definition of Φ^0. Using this Φ^0, we shall next define the required symplectic transformation ψ in (35) implicitly by means of a generating function $\Sigma(\omega T, h)$, which is quasiperiodic in T,

$$\psi : \begin{array}{rcl} \tau & = & T + \Sigma_h(\omega T, h), \\ K & = & h + \Sigma_T(\omega T, h). \end{array} \tag{40}$$

In order to achieve our aim (36) we have to solve the following equation for Σ:

$$I^0(\omega T, h + \Sigma_T(\omega T, h)) = \Phi^0(h), \tag{41}$$

where the dependence on the parameter ε is again neglected. In view of (37) equation (41) is equivalent to

$$h + \Sigma_T(\omega T, h) = K^0(\omega T, \Phi^0(h)).$$

Therefore, the function $\Sigma(\xi, h)$ solves the following partial differential equation on T^N having the constant coefficients $(\omega_1, \ldots, \omega_N) = \omega$:

$$\sum_{j=1}^{N} \omega_j \frac{\partial}{\partial \xi_j} \Sigma(\xi, h) = K^0(\xi, \Phi^0(h)) - h. \tag{42}$$

Since, by our assumption, the frequencies ω satisfy the Diophantine conditions (5) and, by construction, the mean value over the torus of the right-hand side of (42) vanishes in view of (39), there is a unique solution $\Sigma(\xi, h)$ periodic in ξ and having vanishing mean value $[\Sigma](h) = 0$. Because of the well-known small divisor phenomenon, however, this solution loses derivatives, so that

$$\Sigma \in C^{k-\tau}(T^N \times D) \tag{43}$$

if the right-hand side is in C^k, where the parameter τ in (44) is the same as in the Diophantine condition (5). This is well known and we refer to [10] and [16] for a proof. In view of

$$1 + \Sigma_{Th}(\omega T, h) = K_h^0(\omega T, \Phi^0(h)) \; \Phi_h^0(h)$$

$$= \frac{\Phi_h^0}{I_K^0} > 0,$$

the relation (40) indeed defines a symplectic transformation ψ of the form (35) . It is of class $C^{k-\tau-1}$; the extra loss of smoothness is a result of the differentiation in (40). Moreover, by the well-known properties of quasiperiodic functions proved in the book by Siegel and Moser on celestial mechanics [17, §36], one readily verifies that the functions u and v in the transformation ψ are quasiperiodic in T still with the same frequencies ω. The same conclusion follows for the functions representing the inverse transformation of ψ.

Recalling that $T = \varepsilon t$, we now replace τ by $\varepsilon\tau$ and arrive, in view of the Hamiltonian equations corresponding to the function (36), at the equations

$$\frac{d\tau}{d\varphi} = \Phi_h \; (\omega\varepsilon\tau, h, \varphi, \varepsilon), \tag{44}$$

$$\frac{dh}{d\varphi} = -\Phi_\tau \; (\omega\varepsilon\tau, h, \varphi, \varepsilon).$$

The Hamiltonian function

$$\Phi(\varepsilon\omega\tau, h, \varphi, \varepsilon) = \Phi^0(h, \varepsilon) + \varepsilon\Phi^1(\varepsilon\omega\tau, h, \varphi, \varepsilon) \tag{45}$$

is quasiperiodic in τ with frequencies $\varepsilon\omega$. Moreover, $\Phi(\xi, h, \varphi, \varepsilon)$ belongs to $C^{k-1-\tau}(T^N \times D \times S^1 \times \mathbf{R})$ and, by construction,

$$\Phi^0(h, \varepsilon) = c h^\gamma + O(\varepsilon) \tag{46}$$

for a constant $c > 0$.

3.5. Back to the angle and action variables. Proceeding as in step (3.3), we next choose the variables $\varphi, I,$ and τ as the new position, momentum, and time variables. These variables then satisfy the Hamiltonian equations whose Hamiltonian function is the inverse function of $h \to \Phi(\varepsilon\omega\tau, h, \varphi, \varepsilon)$ denoted by $h(\varphi, I, \varepsilon\omega\tau, \varepsilon)$ and thus satisfying

$$\Phi\left(\varepsilon\omega\tau, h(\varphi, I, \varepsilon\omega\tau, \varepsilon)\right) = I.$$

Therefore, the Hamiltonian equations become, on the extended phase space $(\varphi, I, \tau) \in S^1 \times \mathbf{R} \times \mathbf{R}$,

$$\frac{d\varphi}{d\tau} = h_I\left(\varphi, I, \varepsilon\omega\tau, \varepsilon\right),$$

$$\frac{dI}{d\tau} = -h_\tau\left(\varphi, I, \varepsilon\omega\tau, \varepsilon\right).$$

The Hamiltonian function h is of class $C^{k-\tau-1}$, depends quasiperiodically on the time τ, and is of the form

$$h\left(\varphi, I, \varepsilon\omega\tau, \varepsilon\right) = h^0\left(I, \varepsilon\right) + \varepsilon h^1\left(\varphi, I, \varepsilon\omega\tau, \varepsilon\right), \tag{47}$$

where h^0 is the inverse function of Φ^0, which thus satisfies

$$h^0\left(I, \varepsilon\right) = cI^\beta + O(\varepsilon^\alpha)$$

with constants $\beta = \frac{2n+2}{n+2}$, $\alpha > 0$, and $c > 0$.

3.6. Transformation into a system near an integrable one. In order to remove the dependence on φ in the $0(\varepsilon)$-terms of the Hamiltonian (46) we seek a time-dependent symplectic transformation $\psi : (\varphi, I) \to (x, y)$ between two annuli given by means of a generating function $S = S\left(\varphi, y, \varepsilon\omega\tau, \varepsilon\right)$, implicitly, via

$$\psi : \quad \begin{aligned} I &= y + \varepsilon S_\varphi\left(\varphi, y\right), \\ x &= \varphi + \varepsilon S_y\left(\varphi, y\right). \end{aligned} \tag{48}$$

Inserting (48) into (47) and expanding in ε leads to

$$\begin{aligned} &h^0\left(y + \varepsilon S_\varphi\right) + \varepsilon h^1\left(x + \varepsilon S_y\, , y + \varepsilon S_\varphi\right) \\ &\quad = h^0\left(y\right) + \varepsilon h_I^0\left(y\right) S_\varphi + \varepsilon h^1\left(x, y\right) + O\left(\varepsilon^2\right). \end{aligned}$$

To kill the angle dependence in the $O\left(\varepsilon\right)$ term above we require

$$h_I^0\left(y\right) S_\varphi + h^1\left(x, y, \varepsilon\omega\tau\right) = \left[h^1\right]\left(y, \varepsilon\omega\tau\right)$$

with the mean value over S^1 defined by

$$\left[h^1\right]\left(y, \varepsilon\omega\tau\right) = \int_0^1 h^1\left(x, y, \varepsilon\omega\tau\right) dx, \tag{49}$$

and we find

$$S\left(\varphi, y, \varepsilon\omega\tau\right) = \frac{1}{h_I^0\left(y\right)} \int_0^\varphi \left\{\left[h^1\right] - h^1\right\} dx,$$

which is periodic in φ and quasiperiodic in τ. For the transformed Hamiltonian function $H(x,y,\varepsilon\omega\tau,\varepsilon) = h\circ\psi + \varepsilon S_\tau$ we therefore conclude

$$H(x,y,\varepsilon\omega\tau,\varepsilon) = h^0(y,\varepsilon) + \varepsilon[h^1](y,\varepsilon\omega\tau,\varepsilon) + \varepsilon^2 h^2(x,y,\varepsilon\omega\tau,\varepsilon). \tag{50}$$

By repeating the same procedure but replacing εS by $\varepsilon^2 S$, of course with a different function S, the Hamiltonian (50) is transformed into a new Hamiltonian of the form

$$H(x,y,\varepsilon\omega\tau,\varepsilon) = h^0(y,\varepsilon) + \varepsilon[h^1](y,\varepsilon\omega\tau,\varepsilon) + \varepsilon^2[h^2](y,\varepsilon\omega\tau,\varepsilon) + \varepsilon^3 h^3(x,y,\varepsilon\omega\tau,\varepsilon). \tag{51}$$

Now, in order to remove the time dependence from the dominant part in (51) consisting of the first three terms, one again carries out step (3.3), then step (3.4), and then step (3.5) and finally arrives at the following time-dependent Hamiltonian function $H(x,y,\varepsilon\omega t,\varepsilon)$, which in action and angle variables $(x,y) \in S^1 \times D$ for some open and bounded interval $D \subset \mathbf{R}^+$, is given by

$$H(x,y,\varepsilon\omega t,\varepsilon) = H_0(y,\varepsilon) + \varepsilon^3 H_1(x,y,\varepsilon\omega t,\varepsilon), \tag{52}$$

with

$$H_0(y,\varepsilon) = cy^\beta + O(\varepsilon^\alpha)$$

for positive constants c, α, and $\beta = \frac{2n+2}{n+2}$. The function H belongs to $C^{k-2\tau-4}$ and, moreover, is quasiperiodic in time t with the frequencies $\varepsilon\omega$. On the domain $(x,y,t) \in S^1 \times D \times \mathbf{R}$ the system described by (52) turns out to be sufficiently near the integrable system, which is described by the Hamiltonian $H_0(y,\varepsilon)$ provided ε is small. This is the content of the next and last step in the proof of Theorem 1.

3.7. Existence of an invariant cylinder, proof of Theorems 1 and 2. We consider the Hamiltonian system (52) in $S^1 \times D$, where D is a bounded interval of the positive real axis, H is periodic in x, and $\xi = (\xi_1,\ldots,\xi_N)$:

$$H(x,y,\xi,\varepsilon) = H_0(y,\varepsilon) + \varepsilon^3 H_1(x,y,\xi,\varepsilon). \tag{53}$$

We are looking for quasiperiodic solutions having the frequencies $(\alpha,\varepsilon\omega) \in \mathbf{R}^{1+N}$, where ω are the prescribed frequencies of Theorem 1. In more geometric terms we look for a differentiable mapping

$$w : T^{1+N} \to S^1 \times D, \tag{54}$$

$w(\theta,\xi) = (u(\theta,\xi), v(\theta,\xi))$, where $u(\theta,\xi) - \theta$ and $v(\theta,\xi)$ are periodic functions in θ and ξ, which maps the constant vector field V on T^{1+N}, given by

$$V: \begin{aligned} \dot\theta &= \alpha, \\ \dot\xi &= \varepsilon\omega \end{aligned} \tag{55}$$

into the given Hamiltonian vector field belonging to (53), thus satisfying

$$D_V w = dw\begin{pmatrix}\alpha\\ \varepsilon\omega\end{pmatrix} = J\nabla H(w(\theta,\xi),\xi) \tag{56}$$

for all $(\theta, \xi) \in T^{1+N}$. Here ∇ stands for the gradient with respect to the variables (x, y) and

$$D_V = \alpha \frac{\partial}{\partial \theta} + \varepsilon \sum_{j=1}^{N} \omega_j \frac{\partial}{\partial \xi j},$$

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{57}$$

From (56) it then follows that a solution $(\theta(t), \xi(t)) = (\alpha t, \varepsilon \omega t)$ of V in (55) is mapped into the quasiperiodic solution $z(t) = w(\alpha t, \varepsilon \omega t)$ of the Hamiltonian system

$$\dot{z}(t) = J \nabla H(z(t), \varepsilon \omega t, \varepsilon). \tag{58}$$

One concludes, in particular, that the cylinder

$$\hat{w} : S^1 \times \mathbf{R} \to S^1 \times D \times \mathbf{R}, \tag{59}$$

defined by $\hat{w}(\theta, t) = (w(\theta, \varepsilon \omega t), t)$, is tangential to $(J \nabla H(x, y, \varepsilon \omega t), 1)$, the Hamiltonian vector field in the phase space $S^1 \times D \times \mathbf{R}$. The solutions on this cylinder are, moreover, quasiperiodic. The required map w is, in view of (56), a solution of the nonlinear partial differential equation

$$D_V w = J \nabla H(w, \xi, \varepsilon). \tag{60}$$

In the special case of the integrable system defined by $H_0(y, \varepsilon)$, which does not depend on the torus variables $(\theta, \xi) \in T^{1+N}$, the solutions w of (60) are simply the injection mappings

$$w : T^{1+N} \to S^1 \times D, \quad w(\theta, \xi) = (\theta, y), \tag{61}$$

where y is determined by the vector field $V = (\alpha, \varepsilon \omega)$ via

$$\alpha = \frac{\partial H_0}{\partial y}(y, \varepsilon). \tag{62}$$

If $\varepsilon > 0$ and small, then the system H is a perturbation of this integrable system and we shall apply a well-known existence statement of Moser [18] in order to guarantee solutions w of (60) nearby.

First we observe that, by construction, the function H in (52) satisfies

$$C < \frac{\partial^2 H_0}{\partial y^2}(y, \varepsilon) < C^{-1}, \quad y \in D \tag{63}$$

for a positive constant $C > 0$, which is independent of ε for small ε. This is the twist condition. Consequently, if the prescribed frequencies $\omega \in \mathbf{R}^N$ satisfy the Diophantine conditions (5) of the theorem, then we find, for every given $\varepsilon > 0$, a point $y \in D$ such that $\alpha = (\partial H_0 / \partial y)(y, \varepsilon)$ satisfies

$$|\alpha \cdot k + \langle \varepsilon \omega, j \rangle| \geq \varepsilon \gamma (|k| + |j|)^{-\tau} \tag{64}$$

for all $(k,j) \in \mathbf{Z}^{1+N}\backslash\{0\}$ with the constants $\gamma > 0$ and $\tau > N$ as in (5). Indeed, one readily verifies, using $\tau > N$, that in every finite interval I the complement of those real numbers α in I which fail the estimates (64) are a set of Lebesgue measure $O(\varepsilon)$. Secondly, the Hamiltonian function H is sufficiently smooth: $H \in C^l$ for $l > 2\tau + 2$. Indeed, from step (3.6) we have, by construction, $H \in C^{k-2\tau-4}$ and, by assumption, $k > 4\tau + 6$. Thirdly, the perturbation is sufficiently small in the sense that

$$\left(\frac{1}{\varepsilon\gamma}\right)^2 \|H - H_0\|_{C^l} = \frac{\varepsilon}{\gamma^2}\|H_1\|_{C^l} = O(\varepsilon). \tag{65}$$

In view of (63), (64), and (65) we can apply Moser's theorem in [18] together with its improvements from Salamon in [20] and Salamon and Zehnder in [16], which remove the analycity requirement for the unperturbed system. We conclude that, for $0 < \varepsilon \leq \varepsilon^*$ small and $\alpha = (\partial H_0/\partial y)(y,\varepsilon)$ satisfying (64), there exists a solution $w = w_\varepsilon$ of (60). Moreover, this solution $w_\varepsilon = (u_\varepsilon, v_\varepsilon)$ satisfies

$$\begin{aligned} \|u_\varepsilon(\theta,\xi) - \theta\|_{C^1} &= O(\varepsilon), \\ \|v_\varepsilon(\theta,\xi) - y\|_{C^1} &= O(\varepsilon), \end{aligned} \tag{66}$$

so that w_ε is indeed close to the map $w = w_0$ in (61) for the integrable system. It belongs to the same α.

Summarizing, for every $\varepsilon > 0$ sufficiently small we have constructed an invariant cylinder (59). Going back to the original coordinates we conclude, in view of the scaling in step 1, that to every initial condition $(x(0), \dot{x}(0)) \in \mathbf{R}^2$ there is an invariant cylinder as described in the introduction containing the corresponding solution $(x(t), \dot{x}(t))$ of equation (1) in its interior, so that $\sup\{x(t)^2 + \dot{x}(t)^2, t \in \mathbf{R}\} < \infty$. This finishes the proof of Theorem 1.

We remark that, in order to apply the above small denominator techniques to our problem at hand, one simply extends the Hamiltonian system by considering the function

$$\hat{H}(x,\xi,y,\eta) = \varepsilon\langle\omega,\eta\rangle + H(x,y,\xi,\varepsilon)$$

on the extended phase space $T^{1+N} \times \mathbf{R}^{1+N}$ with its standard symplectic structure. The integrable part of $\hat{H}$ is then given by

$$\hat{H}_0(y,\eta,\varepsilon) = \varepsilon\langle\omega,\eta\rangle + H_0(y,\varepsilon).$$

The distinguished invariant torus of this integrable system, defined by $T^{1+N} \times \{y,\eta\}$ for $\alpha = (\partial H_0/\partial y)(y,\varepsilon)$ and $\eta = 0$, which has the frequencies $(\alpha, \varepsilon\omega)$, is then continued under the perturbation. For the existence proof of this continuation one simply applies the standard transformation technique, restricting, however, the symplectic transformations used to the subgroup of those transformations leaving the ξ variables fixed.

4. Proof of Theorem 3 (on general potentials). In this section we shall sketch the proof of Theorem 3. The many tedious technical details are the same as in [15] and the proof of Theorem 1 and will be omitted. We first carry out the formal steps which put $\ddot{x} + V_x(x,\omega t) = 0$ into a suitable normal form in the region $x^2 + \dot{x}^2 \geq C$ in $\mathbf{R}^2$ for $C > 0$ large. Afterwards we follow up with the estimates.

4.1. The formal normal form. In contrast to the polynomial case of Theorem 1 we do not rescale until later.

Step 1. We first introduce the action and angle variables $(x, \dot{x}, t) \to (\theta, I, t)$ by freezing a value t and assigning to (x, y, t) the triple (θ, I, t) as follows:

$$\begin{aligned} \theta &= S_I\,(x, I, \omega t), \\ y &= S_x\,(x, I, \omega t), \end{aligned} \tag{67}$$

where

$$S\,(x, I, \omega t) \;=\; \int_0^x y dx$$

is the integral taken along the level curve $\frac{1}{2}\,y^2 + V\,(x, \omega t) = \text{const}$, which encloses the area I in the (x, y)-plane. The resulting Hamiltonian in the (θ, I, t) variables is then

$$H\,(\theta, I, \omega t) \;=\; H_0\,(I, \omega t) \;+\; H_1\,(\theta, I, \omega t) \tag{68}$$

with

$$H_1 \;=\; S_t \;=\; \omega \cdot \frac{\partial}{\partial \xi} S\,(\theta, I, \omega t).$$

Step 2. Now, proceeding as in (3.3) we choose t, H, θ and $I = I\,(\omega t, H, \theta)$ as the new position, momentum, time, and Hamiltonian function, respectively. Here $I\,(\omega t, H, \theta)$ is the inverse function of $I \mapsto H\,(\theta, I, \omega t)$. Defining $I_0\,(\omega t, H)$ as the inverse of $I \mapsto H_0\,(I, \omega t)$ we rewrite the Hamiltonian describing the system in the form

$$I\,(\omega t, H, \theta) \;=\; I_0\,(\omega t, H) \;+\; I_1\,(\omega t, H, \theta), \tag{69}$$

thus defining I_1 with θ playing the role of the time.

Step 3. Removing the t-dependence in the "leading" term H_0, following §3.4 we arrive at the equivalent Hamiltonian system defined by

$$J\,(\omega \tau, h, \theta) \;=\; J_0\,(h) \;+\; J_1\,(\omega \tau, h, \theta). \tag{70}$$

Step 4. Proceeding as in §3.5, we go back to the variables (θ, J, τ) as the new position, momentum, and time and arrive at the new system with the Hamiltonian

$$h\,(\theta, J, \omega \tau) \;=\; h_0\,(J) \;+\; h_1\,(\theta, J, \omega \tau), \tag{71}$$

where h is the inverse function of $h \mapsto J\,(\omega \tau, h, \theta)$ and h_0 is the inverse of $h \mapsto J_0\,(h)$.

Step 5. In order to work on a bounded interval for the action variable we rescale (θ, J) into (x, y) by setting

$$x \;=\; \theta, \qquad y \;=\; \varepsilon J, \tag{72}$$

where $\varepsilon > 0$ is small and y varies in a bounded interval $D \subset \mathbf{R}^+$. The new variables satisfy the Hamiltonian system corresponding to the Hamiltonian function

$$\varepsilon\, h\,(x, y, \omega \tau) \;=\; \varepsilon\, h_0\left(\frac{y}{\varepsilon}\right) + \varepsilon\, h_1\left(x, \frac{y}{\varepsilon}, \omega \tau\right). \tag{73}$$

Rescaling the time by

$$\tau = \frac{t}{\varepsilon h_0(\frac{1}{\varepsilon})} \tag{74}$$

and introducing the abbreviation $T(\varepsilon) = \varepsilon h_0(\frac{1}{\varepsilon})$, we set

$$\Omega = \Omega(\varepsilon) = \frac{\omega}{T(\varepsilon)} \tag{75}$$

and arrive at the new Hamiltonian system given by

$$\hat{h}(x, y, \Omega t, \varepsilon) = \hat{h_0}(y, \varepsilon) + \varepsilon^\beta \hat{h_1}(x, y, \Omega t, \varepsilon) \tag{76}$$

with the functions $\hat{h_0}$ and $\hat{h_1}$ defined by

$$\hat{h_0}(y, \varepsilon) = \frac{1}{h_0(\frac{1}{\varepsilon})} h_0\left(\frac{y}{\varepsilon}\right),$$

$$\hat{h_1}(x, y, \Omega t, \varepsilon) = \frac{1}{\varepsilon^\beta h_0(\frac{1}{\varepsilon})} h_1\left(x, \frac{y}{\varepsilon}, \Omega t\right).$$

The constant $\beta > 0$ will be chosen later such that $\hat{h_1}$, together with all its finitely many derivatives, is bounded independently of $\varepsilon > 0$.

Step 6. After K further symplectic transformations as in §3.6, we find the following normal form for the Hamiltonian function on an annulus A with coordinates x and y:

$$H(x, y, \Omega t, \varepsilon) = H_0(y, \varepsilon) + \varepsilon^{K\beta} H_K(x, y, \Omega t, \varepsilon), \tag{77}$$

where H is periodic in x and quasiperiodic in t with the frequencies $\Omega = \Omega(\varepsilon)$ as defined in (75). All the transformations result in the loss of $2\tau + 3 + K$ derivatives.

4.2. Estimates. In order to apply the theorems of existence of quasiperiodic solutions as in §3.7 for the flow of the Hamiltonian system given by H in (77), we need the following conditions (i)–(iii) to be satisfied:

(i) There exist constants $0 < C_1 < C_2$ independent of $\varepsilon > 0$ such that

$$C_1 \leq \left(\frac{\partial}{\partial y}\right)^2 H_0(y, \varepsilon) \leq C_2 \tag{78}$$

for all $y \in D$, where D is a bounded interval in $\mathbf{R}^+$.

(ii) For every $\varepsilon > 0$ there is a $y \in D$ such that $\alpha = \frac{\partial}{\partial y} H_0(y, \varepsilon)$ satisfies the Diophantine conditions

$$|\alpha k + \langle \Omega, j \rangle| \geq \gamma |\Omega| \quad (|k| + |j|)^{-\tau} \tag{79}$$

for all $(k, j) \in \mathbf{Z} \times \mathbf{Z}^N \setminus \{0\}$, where $\Omega = \Omega(\varepsilon)$ is given by (75).

(iii) The perturbation satisfies

$$\left(\frac{1}{\gamma|\Omega|}\right)^2 \|\varepsilon^{K\beta} H_K\|_{C^l} = O(\varepsilon^\sigma) \tag{80}$$

for some $l > 2\tau + 2$ and all $\varepsilon > 0$ with a constant $\sigma > 0$.

The following two propositions will allow the verification of (78)–(80). The first proposition is a restatement of Theorems 3 and 4 in [15, p. 47].

PROPOSITION 1. *If the potential $V(x,\xi)$ satisfies the assumptions of Theorem 3 with $|\alpha|+|k| \le d$ then the Hamiltonian function (68) satisfies for I large the following estimates: There are positive constants $C_{k,\alpha}$ and $\delta = \delta(\mu, b, d)$ such that*

$$|\partial_\xi^\alpha \partial_I^k H_1(\theta, I, \xi)| \le C_{k,\alpha}\, I^{-k-\delta}\, H_0(I,\xi), \quad |\alpha| + |k| \le d-1. \tag{81}$$

Moreover, the function $I_0(\xi, H)$ in (69) satisfies for large H,

$$|\partial_\xi^\alpha \partial_H^k I_0(\xi, H)| \le C_{k,\alpha}\, H^{-k}\, I_0(\xi, H), \quad |\alpha| + |k| \le d-1 \tag{82}$$

and for $0 \le k \le 2$,

$$H^{-k}\, I_0(\xi, H) \le C\, |\partial_H^k I_0(\xi, H)| \tag{83}$$

for a constant $C > 0$.

The next proposition is a restatement of Theorem 5.1 in [15].

PROPOSITION 2. *If estimates (81)–(83) hold true then the Hamiltonian function (71) denoted by $J(\xi, h, \theta) = J_0(h) + J_1(\xi, h, \theta)$ satisfies the following estimates for h large: There are positive constants $C, C_k, C_{k,\alpha}, a, a_1$, and σ, such that*

$$\frac{1}{C}\, h^{-a} \le \left| \frac{\partial}{\partial h} J_0(h) \right| \quad \text{and} \quad \frac{1}{C}\, h^{a_1} \le J_0(h), \tag{84}$$

$$\frac{1}{C}\, h^{-k}\, J_0(h) \le \left| \left(\frac{\partial}{\partial h} \right)^k J_0(h) \right| \quad \text{for } 0 \le k \le 2, \tag{85}$$

$$\left| \left(\frac{\partial}{\partial h} \right)^k J_0(h) \right| \le C_k\, h^{-k}\, |J_0(h)|, \quad |k| \le d-2, \tag{86}$$

$$|\partial_\xi^\alpha \partial_h^k J_1(\xi, h, \theta)| \le C_{k,\alpha}\, h^{-k}\, J_0(h)^{1-\sigma}, \quad |\alpha| + |k| \le d-2. \tag{87}$$

By using the techniques developed in [15], one can follow these estimates through all our coordiante transformations in order to verify (78) and (80). In particular, estimates (84)–(87) imply the existence of constants $\beta = \beta(b, \mu, d)$ and $\mu_0 = \mu_0(b)$ such that

$$\|H_K\|_{C^{d-2\tau-3-K}} = O(1).$$

Furthermore, there exists $\alpha = \alpha(b) > 0$ such that

$$\varepsilon h_0 \left(\frac{1}{\varepsilon} \right) = O(\varepsilon^{-\alpha}).$$

To verify (80) we combine the last two estimates

$$\left(\frac{1}{\gamma |\Omega|} \right)^2 \|\varepsilon^K H_K\|_{C^{d-2\tau-3-K}} = O\left(\varepsilon^{K\beta - 2\alpha}\right),$$

and it remains to show that $K\beta(b,\mu,d) - 2\alpha > 0$ and $d - 2\tau - 3 - K > 2\tau + 2$ hold for some $K > 0$ or, equivalently,

$$2\frac{\alpha(b)}{\beta(b,\mu,d)} < K < d - 4\tau - 5.$$

For such K to exist it suffices to verify that

$$d - 4\tau - 5\frac{2\alpha(b)}{\beta(b,\mu,d)} \geq 1 \quad \text{and} \quad \beta(b,\mu,d) > 0. \tag{88}$$

First, $\beta(b,0,d) = \beta(b) > 0$ is d-independent, according to the arguments in [15], and $\beta(b,\mu,d)$ can be chosen as continuous in μ for $\mu \geq 0$. Thus, choosing

$$d \;=\; 4\tau + 7 + 2\frac{\alpha(b)}{\beta(b)}$$

guarantees that (88) holds for all $\mu \in [0,\mu_0(b)]$, where $\mu_0(b) > 0$ can be estimated explicitly. With this choice of d there exists K with which the smallness condition (80) is satisfied. This finishes the sketch of the proof of Theorem 3.

5. Proof of Theorem 4 (unbounded solutions). The idea is to create $p(t)$, giving a particular solution $y(t)$ of (11) a "helping kick" to the right each time the solution passes through the interval $-1 \leq x \leq 1$ from -1 to $+1$, and make $p(t) = 0$ at all other times. With such a $p(t)$ the energy along the solution will increase during each passage from -1 to $+1$ while remaining constant between consecutive passages. If the "kicks" do not weaken too much with the number of passages, the errors will grow without bound. The precise construction is as follows: We first consider an auxiliary nonconservative system

$$\ddot{y} + y^3 = f(y,\dot{y}) \tag{89}$$

for a smooth function $f \in C^\infty(\mathbf{R}^2)$ defined by

$$f(y,\dot{y}) = h(y)\cdot\frac{g(\dot{y})}{\dot{y}^r}, \tag{90}$$

where $r \in \mathbf{N}$ is as it was in the statement of the theorem and h and $g \in C^\infty(\mathbf{R})$ satisfy

$$h\,(y)\begin{cases} = 0 & \text{if} \quad |y| \geq 1, \\ > 0 & \text{if} \quad |y| < 1, \end{cases}$$

$$g\,(\dot{y})\begin{cases} = 0 & \text{if} \quad \dot{y} \leq 0, \\ > 0 & \text{if} \quad 0 < \dot{y} < 1, \\ = 1 & \text{if} \quad \dot{y} \geq 1. \end{cases} \tag{91}$$

Now, if $y(t)$ is a fixed unbounded solution of (89), we can define $p\,(t) = f\,(y\,(t),\dot{y}\,(t))$ and consider equation (11) with this forcing term p. The function $y\,(t)$ then solves both the conservative equation (11) and the nonconservative equation (89). It turns out that all nontrivial solutions of the latter are unbounded.

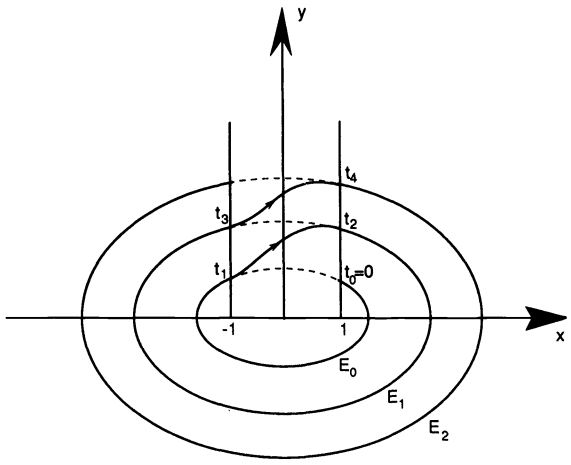


FIG. 5.1.

The energy

$$E(t) := \frac{1}{2}\dot{y}(t)^2 + \frac{1}{4}y(t)^4 \tag{92}$$

along a given solution $y(t)$ of (89) satisfies

$$\frac{d}{dt}E(t) = \dot{y}\, f(y, \dot{y}), \tag{93}$$

which is > 0 if $\dot{y} > 0$ and $|y| < 1$ and 0 otherwise.

To be specific, in the following we consider the solution $y(t)$ with initial values $y(0) = \dot{y}(0) = 1$. Referring to Fig. 5.1, the motion of $(y, \dot{y})$ in the $\mathbf{R}^2$-plane follows the energy curve in $\mathbf{R}^2$ given by $\frac{1}{2}\dot{y}^2 + \frac{1}{4}y^4 = \frac{3}{4}$ in the clockwise direction until the point $(-1, 1)$ is reached at some time $t_1 > t_0 = 0$. At some later moment $t_2 > t_1$ the point $(y, \dot{y})$ will cross the right boundary $\{y = 1, \dot{y} > 1\}$ of the strip. Indeed, this follows by comparison with the equation $\ddot{y} + y^3 = 0$ from the fact that $f(y, \dot{y}) > 0$ in the strip. Furthermore, $E(t_2) > E(t_1)$ by (93), so that the point will start moving along a larger energy curve after having crossed the strip.

Denoting the sequence of consecutive crossings of the vertical boundaries $\{y = \pm 1, \dot{y} > 0\}$ by $0 = t_0 < t_1 < t_2 < \cdots$, and the sequence of corresponding energy values between consecutive crossings by

$$E_n = E(t), \quad t_{2n} \le t \le t_{2n+1},$$

$n = 0, 1, 2, \ldots$, we claim that

$$\lim_{n\to\infty} E_n = \lim_{t\to\infty} E(t) = \infty. \tag{94}$$

Indeed, integration of (93) gives

$$\begin{aligned} E_{n+1} - E_n &= \int_{t_{2n+1}}^{t_{2n+2}} \dot{y} f(y, \dot{y}) dt \\ &= \int_{t_{2n+1}}^{t_{2n+2}} h(y) \frac{1}{\dot{y}^{r-1}} dt = \int_{-1}^{1} h(y) \cdot \frac{1}{\dot{y}^r} dy. \end{aligned} \tag{95}$$

The inequalities

$$E_n < E(t) < E_{n+1} \text{ for } t_{2n+1} < t < t_{2n+2}$$

result, in view of $|y(t)| < 1$ and $E_n > \frac{1}{2}$, in

$$\sqrt{E_n} \leq \dot{y}(t) \leq \sqrt{2}\sqrt{E_{n+1}} \quad \text{for } t_{2n+1} \leq t \leq t_{2n+2}. \tag{96}$$

From (95) and (96) we conclude

$$\frac{a}{E_{n+1}^{r/2}} < E_{n+1} - E_n < \frac{b}{E_n^{r/2}}, \quad n = 0, 1, 2, \ldots \tag{97}$$

for two constants $0 < a < b$. Consequently, E_n is a strictly monotone increasing sequence; moreover, $\lim E_n = \infty$. Indeed, if $\lim E_n = E_\infty < \infty$ then taking the limit in (97) leads to the contradiction $0 < a/E_\infty^{r/2} = 0$. We have proved claim (94) and conclude, in view of (92), that the solution $y(t)$ of (89) is unbounded.

In order to prove claim (10) in Theorem 4 for $p(t) = f(y(t), \dot{y}(t))$, we note first that, for $t_{2n+1} \leq t \leq t_{2n+2}$, we have $\dot{y}(t) > 0$ and thus

$$p(t) = h(y)\dot{y}^{-r},$$

while $p(t) = 0$ otherwise. Since $\lim E_n = \infty$, estimate (96) implies $\lim_{t\to\infty} p(t) = 0$.

Similarly, differentiating p and observing that $y(t)$ satisfies equation (89), one readily verifies the estimate

$$|D^j p\ (t)| \leq C(h)\frac{1}{\dot{y}(t)^{r-j}}, \quad \text{if } t_{2n+1} < t < t_{2n+2}, \tag{98}$$

while $D^j p(t) = 0$ otherwise, so that $\lim_{t\to\infty} D^j p(t) = 0$ if $1 \leq j \leq r-1$. Here, the constant $C(h)$ depends only on h and its derivatives and, moreover, $C(\varepsilon h) \leq \varepsilon C_1(h)$ for ε small. Replacing $h(y)$ by $\varepsilon h(y)$, we get the required estimate (10) for the forcing $p(t)$.

Finally, we prove estimates (12) and (13). Since, by (94) and (97), $c_1 < E_{n+1}/E_n < c_2$ for two positive constants $c_1 < c_2$, we have

$$\frac{a}{E_n^{r/2}} < E_{n+1} - E_n < \frac{b}{E_n^{r/2}} \tag{99}$$

for two constants $0 < a < b$, which are different from the previous constants. To estimate E_n we define the continuous function $\eta(t)$ by linearly interpolating E_n:

$$\eta(t) := (t-n)E_{n+1} + (1-t+n)E_n$$

if $n \leq t \leq n+1$, so that $E_n = \eta(n)$. Thus $\eta(t)$ is a strictly monotone increasing function whose right derivative D_r satisfies, in view of (99),

$$\frac{a}{\eta(t)^{r/2}} \leq D_r\eta(t) \leq \frac{b}{\eta(t)^{r/2}} \tag{100}$$

for all $t \geq 0$; this is the differential inequality which interpolates (99). Comparison with the solutions $\xi_a(t)$ and $\xi_b(t)$ of the two equations

$$\dot{\xi} = \frac{\alpha}{\xi^{r/2}} \quad \text{with } \alpha = a \text{ and } \alpha = b,$$

with initial conditions $\xi_a(0) = \eta(0) = \xi_b(0)$, leads to $\xi_a(t) \le \eta(t) \le \xi_b(t)$ for $t \ge 0$. Consequently, there are constants $0 < A < B$ such that

$$A\, n^{\frac{1}{1+r/2}} \le E_n \le B\, n^{\frac{1}{1+r/2}} \tag{101}$$

for $n = 0, 1, 2, \ldots$. In order to relate n with t_n we note that

$$T(E_{n+1}) < t_{2n+2} - t_{2n} < T(E_n), \tag{102}$$

where $T(E)$ denotes the period of the solutions of $\ddot{x} + x^3 = 0$ with energy $E = \frac{1}{2}\dot{x}^2 + \frac{1}{4}x^4$. A simple calculation gives $T(E) = \tau E^{-1/4}$ for some constant $\tau > 0$. Using this in (102), we conclude from (101) that

$$\frac{a}{n^{\frac{1}{2r+4}}} < t_{2n+2} - t_{2n} < \frac{b}{n^{\frac{1}{2r+4}}} \tag{103}$$

for two constants $0 < a < b$, which are different from the ones in previous formulas. Adding up inequalities (103), we obtain the estimate

$$c\, n^{\frac{2r+3}{2r+4}} < t_n < C\, n^{\frac{2r+3}{2r+4}} \tag{104}$$

for two constants $0 < c < C$. Recalling the definition of $E(t)$, we conclude from (101) and (104) that

$$a\, t^{\frac{4}{2r+3}} < E(t) < b\, t^{\frac{4}{2r+3}},\ t \ge 1 \tag{105}$$

for two constants $0 < a < b$. Finally, in view of (96) and (98) we now conclude

$$|D^j p(t)| \le c_1 \frac{1}{\sqrt{E(t)}^{r-j}} \le c_2 \frac{1}{t^{\frac{2(r-j)}{2r+3}}}$$

as claimed in the theorem. This finishes the proof of Theorem 4.

Acknowledgments. We would like to thank Jürgen Moser for valuable discussions and Claudia Flepp for deciphering and typing a chaotic manuscript.

REFERENCES

[1] J. E. LITTLEWOOD, *Some problems in real and complex analysis*, Heath, Lexington, MA, 1968.
[2] ———, *Unbounded solutions of $y - g(y) = p(t)$, with $p(t)$ periodic and bounded $g(y)/(y) \to \infty$ as $y \to \pm\infty$*, J. London Math. Soc., 41 (1966), pp. 491–507.
[3] M. LEVI, *On Littlewood's counterexample of unbounded motions in superquadratic potentials*, Dynam. Report. (N.S.), 1 (1992), pp. 113–124.
[4] Y. LONG, *Unbounded solution of a superlinear Duffing's equation*, Nankai University, Tianjin, 1989, preprint.
[5] G. R. MORRIS, *A case of boundedness in Littlewood's problem on oscillatory differential equations*, Bull. Austral. Math. Soc., 14 (1976), pp. 71–93.
[6] S. LAEDERICH AND M. LEVI, *Invariant curves and time-dependent potentials*, Ergodic Theory Dynamical Systems, 11 (1991), pp. 365–378.
[7] R. DIECKERHOFF AND E. ZEHNDER, *Boundedness of solutions via the twist theorem*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 14 (1987), pp. 79–95.
[8] J. W. NORRIS, *Boundedness in periodically forced second order conservative systems*, J. London Math. Soc. (2), 45 (1992), pp. 97–112.
[9] L. BIN, *Boundedness for solutions of nonlinear Hill's equations with Periodic Forcing Terms via Moser's twist Theorem*, J. Differential Equations, 79 (1989), pp. 304–315.

[10] H. RÜSSMANN, *On the existence of invariant curves of twist mappings of an annulus*, in Geometric Dynamics, Springer Lecture Notes in Math. 1007, J. Palis, ed., Springer-Verlag, Berlin, New York, Heidelberg, Tokyo, 1981, pp. 677–718.

[11] M. R. HERMAN, *Sur les Courbes Invariantes par les Diffeomorphismes de l'Anneau, Vol.* 1, Astérisque, 103-104 (1983), pp. 1–221, *Vol.* 2, Asterisque, 149 (1986), pp. 1–248.

[12] M. LEVI, *KAM-theory for particles in periodic potentials*, Ergodic Theory Dynamical Systems, 10 (1990), pp. 777–785.

[13] J. MOSER, *Quasi-periodic solutions of nonlinear elliptic partial differential equations*, Bol. Soc. Brasil Mat., 20 (1989), pp. 29–45.

[14] L. CHIERCHIA AND E. ZEHNDER, *Asymptotic Expansions of Quasiperiodic Solutions*, Ann. Scuola. Norm. Sup. Pisa Cl. Sci. (4), 16 (1989), pp. 245–258.

[15] M. LEVI, *Quasiperiodic motions in superquadratic time-periodic potentials*, Comm. Math. Phys., 143 (1991), pp. 43–83.

[16] D. SALAMON AND E. ZEHNDER, *KAM theory in configuration space*, Comm. Math. Helv., 64 (1989), pp. 84–132.

[17] C. L. SIEGEL AND J. K. MOSER, *Lectures on Celestial Mechanics*, Springer-Verlag, New York, Berlin, Heidelberg, 1971.

[18] J. MOSER, *On the construction of almost periodic solutions for ordinary differential equations*, Proc. Internat Conference Functional Analysis and Related Topics, University of Tokyo Press, Tokyo, Japan, 1970, pp. 60–67.

[19] ———, *On invariant curves of area preserving mappings of an annulus Nachr. Akad. Wiss.* Göttingen Math. Phys. Kl. II, 1962, pp 1–20.

[20] D. SALAMON, *The Kolmogorov–Arnold–Moser theorem*, FIM of the ETH, Zürich, 1986, preprint.

[21] J. MATHER, *Destruction of invariant circles*, in Ergodic Theory Dynamical Systems, 8* (1988), pp. 199–214.

[22] C. V. COFFMAN AND D. F. ULLRICH, *On the continuation of solutions of certain nonlinear differential equations*, Monatsh. Math., 71 (1967), pp. 385–392.

[23] V. I. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, Berlin, 1978.

[24] ———, *On the behavior of an adiabatic invariant under a slow periodic change of the Hamiltonian*, Soviet Math. Dokl., 3 (1962), pp. 136–139.

[25] N. N. NEKHOROSHEV, *Exponential estimate of the stability time of near-integrable Hamiltonian systems*, Russian Math. Surveys, 32 N.6. (1977), pp. 1–65.

[26] A. GIORGILLI AND E. ZEHNDER, *Exponential stability for time dependent potentials*, Z. Angew. Math. Phys., to appear.

[27] R. DOUADY, *Applications du Theoreme des Tores invariants*, Ph.D. thesis, University of Paris VII, 1982, pp. III-17–III-21.

[28] P. LOCHAK, *Canonical perturbation theory via simultaneous approximation*, Uspekhi Mat. Nauk, 47 (1992), pp. 59–40.

[29] J. PÖSCHEL, *Nekhoroshev estimates for quasi-convex Hamiltonian systems*, Math. Z., 213 (1993), pp. 187–216.

[30] V. ZHARNITSKY, *Breakdown of stability of motion in superquadratic potentials*, manuscript.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1257–1283, September 1995

010

PERIODIC SOLUTIONS OF A SYSTEM OF COUPLED OSCILLATORS NEAR RESONANCE*

CARMEN CHICONE†

Abstract. A system of autonomous ordinary differential equations depending on a small parameter is considered such that the unperturbed system has an invariant manifold of periodic solutions that is not normally hyperbolic but is normally nondegenerate. The bifurcation function whose zeros are the bifurcation points for families of perturbed periodic solutions is determined. This result is applied to find the periodic solutions near resonance for a two-degrees-of-freedom mechanical system modeling a rotor interacting with an elastic support.

Key words. coupled oscillator, resonance, normal nondegeneracy

AMS subject classifications. 58F14, 58F22, 58F30, 34C15, 34C25

1. Introduction. In this paper we describe an application of the results in [6] to the bifurcation of periodic solutions in a smooth system of coupled oscillators E_ϵ given by

$$\dot{x}_1 = f_1(x_1) + \epsilon g_1(x_1, \dot{x}_1, x_2, \dot{x}_2),$$
$$\dot{x}_2 = f_2(x_2) + \epsilon g_2(x_1, \dot{x}_1, x_2, \dot{x}_2),$$

where $x_i \in \mathbb{R}^2$, $i = 1, 2$, and $\epsilon \in \mathbb{R}$ when the unperturbed system E_0 satisfies the following conditions:

1. The plane autonomous system $\dot{x}_1 = f_1(x_1)$ has an invariant annulus A consisting of periodic solutions (a period annulus) and every periodic solution in A has the same period $\eta_1 > 0$. Such a period annulus is called isochronous with period η.
2. The plane autonomous system $\dot{x}_2 = f_2(x_2)$ has a periodic trajectory Γ with period $\eta_2 > 0$ such that either Γ is a hyperbolic limit cycle or Γ belongs to a period annulus and the derivative of an associated period function at Γ does not vanish.
3. There are relatively prime positive integers K_1 and K_2 such that $K_1\eta_1 = K_2\eta_2$. In this case we say that the periodic trajectory Γ is in resonance with the period annulus A.

A few comments are in order on the conditions just stated. The prime example of an isochronous period annulus is a period annulus of a linear system. However, given any period annulus and any Poincaré section at a point in the period annulus, there is an associated period function that assigns to each point on the section the time of first return to the section. It is easy to see that the requirement of a nonzero derivative of a period function as in (2) above is independent of the choice of section and the point chosen on the periodic trajectory. The hypotheses ensure that $A \times \Gamma$ is an invariant submanifold of the state space for the unperturbed system E_0 of a special type we call a normally nondegenerate period manifold. The condition of normal nondegeneracy defined precisely in §2 ensures that the first-order bifurcation theory in [6] can be applied and the existence of periodic solutions for the perturbed system near the period manifold can be generically determined by computing the simple

* Received by the editors January 27, 1993; accepted for publication December 15, 1993. This research was supported by the Air Force Office of Scientific Research and National Science Foundation grant DMS-9022621.

† Department of Mathematics, University of Missouri, Columbia, Missouri 65211.

zeros of a certain bifurcation function also defined in §2. Of particular interest here is the fact that the period manifold for E_0 is not normally hyperbolic. Thus, while the period manifold usually does not persist after perturbation, some of the periodic solutions on the period manifold can persist. The bifurcation function determines the number and the position of these persistent periodic solutions. In this way entrainment phenomena can be studied for perturbations of systems which do not already contain stable periodic solutions. For background material on bifurcation problems of this type in addition to [5], [6], the following references and their bibliographies are suggested: [1]–[3], [7], [9]–[19].

While higher-dimensional systems can be studied by the same methods, the four-dimensional system E_ϵ illustrates the important features of the general theory and is sufficiently general to have many interesting specializations to physical applications. In §3 we apply the theory to an ubiquitous system of differential equations which we interpret, as in [16], as a model for a rotor interacting with an elastic support. We show the existence of a normally nondegenerate period manifold in the case in which the unperturbed system is weakly nonlinear and also in the fully nonlinear case which corresponds to the rotor strongly influenced by a gravitational field. In both cases the bifurcation function is computed explicitly and the existence of periodic solutions relative to the choice of parameters is determined. These results are augmented by some numerical evidence suggesting the role of these bifurcating families of periodic solutions in determining the global behavior of the perturbed system.

The plan of the paper is as follows: In §2 we review the general theory of [6]. In §3 we specialize the general theory to the case represented by E_ϵ and identify the bifurcation function. These results are applied in §4 to the mechanical system modeling the rotor with elastic support. There, the bifurcation function is computed explicitly in terms of elliptic functions and its zeros are computed. This determines the perturbed periodic solutions of the coupled mechanical oscillators near resonance. In addition, §4 contains a discussion of some numerical experiments that suggest the coexistence, for certain choices of the parameters, of perturbed periodic attractors, as predicted by the bifurcation analysis, and more complicated nonperiodic attractors.

2. Bifurcation theory. In this section, we outline for completeness a result in [6] which will be used in the analysis of the system E_ϵ defined in the introduction. The analysis begins with a smooth system of differential equations F_ϵ given by

$$\dot{x} = f(x) + \epsilon g(x, \dot{x}, \epsilon), \quad x \in \mathbb{R}^{n+1}, \quad \epsilon \in \mathbb{R},$$

where the unperturbed system F_0 contains a normally nondegenerate period manifold. Here, a period manifold $\mathcal{A}$ is a smooth invariant connected $(k+1)$-dimensional submanifold of $\mathbb{R}^{n+1}$ consisting entirely of periodic solutions of the unperturbed system with the additional property that the Poincaré map P associated with any Poincaré section Σ is the identity on $\mathcal{A} \cap \Sigma$. Of course, period manifolds generalize the concept of a period annulus to many dimensions. To define the concept of normal nondegeneracy we need a few more definitions. On a fixed Poincaré section Σ_0, which has nonempty intersection with $\mathcal{A}$, there is some $\epsilon_0 > 0$ and some subsection $\Sigma \subseteq \Sigma_0$ such that the parametrized Poincaré map $P : \Sigma \times (-\epsilon_0, \epsilon_0) \to \Sigma_0$ is given by $(\xi, \epsilon) \mapsto P(\xi, \epsilon)$, where $P(\xi, \epsilon)$ denotes the first return to Σ_0 of the perturbed solution starting at $\xi \in \Sigma$. After choosing coordinates on Σ given by $s : \mathbb{R}^n \to \Sigma$, the parameterized Poincaré map is identified with its local representation $p : \mathbb{R}^n \times (-\epsilon_0, \epsilon_0) \to \mathbb{R}^n$ given by $p(y, \epsilon) := s^{-1} P(s(y), \epsilon)$. This, in turn, allows us to define the parametrized displacement function $\delta : \mathbb{R}^n \times (-\epsilon_0, \epsilon_0) \to \mathbb{R}^n$ by $\delta(y, \epsilon) := p(y, \epsilon) - y$. Now, for $y_* \in \mathbb{R}^n$

such that $s(y_*) \in \Sigma \cap \mathcal{A}$, it is clear that the derivative of the map $y \mapsto \delta(y,0)$, which we denote by $D\delta(y,0)$, when evaluated at y_*, will have a nontrivial kernel containing the tangent space of $\mathcal{A}$. More precisely, if $v \in \mathbb{R}^n$ and $Ds(y_*)v \in T_{s(y_*)}\Sigma \cap T_{s(y_*)}\mathcal{A}$, then $D\delta(y_*,0)v = 0$. Since $\Sigma \cap \mathcal{A}$ is k-dimensional, the kernel of $D\delta(y_*,0)$ has dimension at least k. If this kernel has dimension k for each y such that $s(y) \in \mathcal{A}$, we say that $\mathcal{A}$ is normally nondegenerate. Perhaps a remark is in order on the definition of displacement. One must exercise caution when defining displacement on the manifold Σ. We have avoided the differential geometry necessary to give an intrinsic definition by introducing local coordinates. However, it should be clear that the zero set of the displacement function, the set corresponding to periodic solutions of F_ϵ, is invariant under change of coordinates.

A goal of the theory in [6] is the identification of a bifurcation function $\mathcal{B}$ defined on $\Sigma \cap \mathcal{A}$ whose simple zeros correspond to the initial values of persistent periodic solutions of the unperturbed system. To construct the bifurcation function, we start with a splitting of the tangent bundle over $\mathbb{R}^{n+1}$ into three subbundles: $\mathcal{E}$ generated by the unperturbed vector field, $\mathcal{E}^{\tan}$ tangent to $\mathcal{A}$ but complementary to $\mathcal{E}$, and $\mathcal{E}^{\mathrm{nor}}$ normal to $\mathcal{A}$. In particular, for $y \in \mathcal{A}$ we have $\mathbb{R}^{n+1} = \mathcal{E}(y) \oplus \mathcal{E}^{\tan}(y) \oplus \mathcal{E}^{\mathrm{nor}}(y)$. Such a splitting always exists, but the last two summands are not unique. Next, we define special coordinates on $\mathbb{R}^{n+1}$ near each point $y \in \Sigma \times \mathcal{A}$ which respect the splitting. For this, we choose $\Delta : \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}^{n-k} \to \mathbb{R}^{n+1}$ given by $(s,\theta,\zeta) \mapsto \Delta(s,\theta,\zeta)$ such that (using subscripted variables to denote partial derivatives)

$$\begin{aligned}
\Delta_s(0,\theta,0) &: \mathbb{R} \to \mathcal{E}(\Delta(0,\theta,0)),\\
\Delta_\theta(0,\theta,0) &: \mathbb{R}^k \to \mathcal{E}^{\tan}(\Delta(0,\theta,0)),\\
\Delta_\zeta(0,\theta,0) &: \mathbb{R}^{n-k} \to \mathcal{E}^{\mathrm{nor}}(\Delta(0,\theta,0)).
\end{aligned}$$

Such coordinates are called adapted to the splitting over $\mathcal{A}$. An associated Poincaré section, again denoted by Σ, is given by the image of the map $(\theta,\zeta) \mapsto \Delta(0,\theta,\zeta)$. In these coordinates the kernel of $D\delta(\Delta(0,\theta,0),0)$ corresponds to $\mathcal{E}^{\tan}(\Delta(0,\theta,0))$ and there is a k-dimensional complement to the range of this derivative in $\mathbb{R}^{n+1}$. After choosing coordinates on the range, the linear projection $H(\theta)$ from the tangent space of $\mathbb{R}^{n+1}$ to this range can be represented as a linear map of the form

$$H(\theta) : \mathcal{E} \oplus \mathcal{E}^{\tan} \oplus \mathcal{E}^{\mathrm{nor}}(\Delta(0,\theta,0)) \to \mathbb{R}^k.$$

Next, let $t \mapsto x(t,\theta)$ denote the solution of F_0 with initial condition $x(0,\theta) = \Delta(0,\theta,0)$ and consider the variational equation along this solution, namely,

$$\dot{W} = Df(x(t,\theta))W.$$

This variational equation has a fundamental matrix solution $t \mapsto \Phi(t,\theta)$ with initial value $\Phi(0,\theta) = I$. There are parametrized linear maps

$$\begin{aligned}
&a(t,\theta) : \mathcal{E}^{\mathrm{nor}}(x(0,\theta)) \to \mathcal{E}^{\tan}(x(t,\theta)), \quad b(t,\theta) : \mathcal{E}^{\mathrm{nor}}(x(0,\theta)) \to \mathcal{E}^{\mathrm{nor}}(x(t,\theta)),\\
&c(t,\theta) : \mathcal{E}^{\tan}(x(0,\theta)) \to \mathcal{E}^{\tan}(x(t,\theta)), \quad d(t,\theta) : \mathcal{E}^{\mathrm{nor}}(x(0,\theta)) \to \mathcal{E}(x(t,\theta)),\\
&e(t,\theta) : \mathcal{E}^{\tan}(x(0,\theta)) \to \mathcal{E}(x(t,\theta))
\end{aligned}$$

such that the block form of $\Phi(t,\theta)$ with respect to the splitting is

$$\Phi(t,\theta) = \begin{pmatrix} 1 & e(t,\theta) & d(t,\theta) \\ 0 & c(t,\theta) & a(t,\theta) \\ 0 & 0 & b(t,\theta) \end{pmatrix},$$

and

$$e(0,\theta)=0,\ \ d(0,\theta)=0,\ \ c(0,\theta)=I,\ \ a(0,\theta)=0,\ \ b(0,\theta)=I.$$

Also, the vector field along the unperturbed solution defined by the perturbation, namely, $G(t,\theta) := g(x(t,\theta),\dot{x}(t,\theta),0)$, has a representation relative to the splitting given by

$$G(t,\theta)=\begin{pmatrix} G^{\mathcal{E}}(t,\theta) \\ G^{\tan}(t,\theta) \\ G^{\mathrm{nor}}(t,\theta) \end{pmatrix}.$$

Here, $G(t,\theta)$ is the derivative of $f(x)+\epsilon g(x,\dot{x},\epsilon)$ with respect to ϵ evaluated at $\epsilon=0$. The bifurcation function for the system F_ϵ adapted to the period manifold $\mathcal{A}$ is the function $\mathcal{B}:\mathbb{R}^k\to\mathbb{R}^k$ defined by

$$\mathcal{B}(\theta)=H(\theta)\begin{pmatrix} 0 \\ \mathcal{N}(\theta) \\ \mathcal{M}(\theta) \end{pmatrix},$$

where

$$\mathcal{M}(\theta) := \int_0^{T(\theta)} b^{-1}(s,\theta)G^{\mathrm{nor}}(s,\theta)\,ds,$$
$$\mathcal{N}(\theta) := \int_0^{T(\theta)} c^{-1}(s,\theta)G^{\tan}(s,\theta) - c^{-1}(s,\theta)a(s,\theta)b^{-1}(s,\theta)G^{\mathrm{nor}}(s,\theta)\,ds,$$

and $T(\theta)$ denotes the time of first return to the Poincaré section for the unperturbed solution $t\mapsto x(t,\theta)$. The following theorem is proved in [6].

THEOREM 2.1. *Suppose F_ϵ, given by*

$$\dot{x}=f(x)+\epsilon g(x,\dot{x},\epsilon),\ \ x\in\mathbb{R}^{n+1},\ \ \epsilon\in\mathbb{R},$$

has a normally nondegenerate period manifold $\mathcal{A}$ with an adapted coordinate system given by $(s,\theta,\zeta)\mapsto\Delta(s,\theta,\zeta)$. If θ_0 is a simple zero of the bifurcation function $\theta\to\mathcal{B}(\theta)$ adapted to $\mathcal{A}$, then there is an $\epsilon_*>0$ and a smooth function $\beta:(-\epsilon_*,\epsilon_*)\to\mathbb{R}^k\times\mathbb{R}^{n-k}$ with $\beta(0)=(\theta_0,0)$ such that $\Delta(0,\beta(\epsilon))$ is the initial value for a periodic solution of F_ϵ.

3. Persistent periodic solutions of the coupled oscillator. In this section we apply the results outlined in §2 to the system E_ϵ defined in the introduction. To do this we must identify the bifurcation function. Other, perhaps simpler examples of the identification procedure are given in [6]. In any case, there are several steps.

Step 1 [definition of the period manifold]. Under assumptions 1–3 listed in the introduction, the unperturbed system E_0 has a three-dimensional period manifold given by $\mathcal{A} := A\times\Gamma$. In fact, every solution of the unperturbed system starting on $\mathcal{A}$ has the same period $T_{\mathcal{A}} := K_1\eta_1$.

Step 2 [adapted coordinates]. For vectors $v=(v_1,v_2)$ and $w=(w_1,w_2)$ in $\mathbb{R}^2$, let $\langle v,w\rangle$ denote the usual inner product, $||v||^2 := \langle v,v\rangle$, $v^\perp := (-v_2,v_1)$, and $v\wedge w := \langle w,v^\perp\rangle$. Using these definitions and the unperturbed vector fields f_1 and f_2 on $\mathbb{R}^2$, we define two smooth vector fields $f_1^\perp$ and $f_2^\perp$ on $\mathbb{R}^2$. Also, we let φ_t^i denote

the flow of $\dot{x}_i = f_i(x_i)$ and ψ_t^i denote the flow of $\dot{x}_i = f_i^{\perp}(x_i)$ for $i = 1, 2$. For each $x = (x_1, x_2)$ in $A \times \Gamma$, we define a splitting over $\mathcal{A}$ by

$$\mathcal{E}(x) = \left[\begin{pmatrix} f_1(x_1) \\ f_2(x_2) \end{pmatrix} \right],$$

$$\mathcal{E}^{\tan}(x) = \left[\begin{pmatrix} ||f_1(x_1)||^{-2} f_1^{\perp}(x_1) \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ f_2(x_2) \end{pmatrix} \right],$$

$$\mathcal{E}^{\mathrm{nor}}(x) = \left[\begin{pmatrix} 0 \\ ||f_2(x_2)||^{-2} f_2^{\perp}(x_2) \end{pmatrix} \right],$$

where the square brackets here and hereafter denote the subspace spanned by the enclosed vectors. This gives

$$T_x\mathcal{A} = \mathcal{E}(x) \oplus \mathcal{E}^{\tan}(x) \oplus \mathcal{E}^{\mathrm{nor}}(x).$$

Next, fix $\xi_1 \in A$ and $\xi_2 \in \Gamma$, and define adapted coordinates $\Delta : \mathbb{R}^4 \to \mathbb{R}^4$ by

$$(s, p, q, \zeta) \mapsto \left(\varphi_s^1(\psi_p^1(\xi_1)), \psi_\zeta^2(\varphi_{s+q}^2(\xi_2))\right).$$

If

$$\Sigma_0 := \left\{\Delta(0, p, q, \zeta) \mid (p, q, \zeta) \in \mathbb{R}^3\right\},$$

then there is some open subset $\Sigma \subseteq \Sigma_0$ that is a three-dimensional Poincaré section for E_0 at (ξ_1, ξ_2).

Step 3 [fundamental matrix of variational equation in adapted coordinates]. We consider the fundamental matrix solution $\Phi(t)$ with initial condition $\Phi(0) = I$ for the variational equation

$$\begin{pmatrix} \dot{w}_1 \\ \dot{w}_2 \end{pmatrix} = \begin{pmatrix} Df_1\left(\varphi_t^1(\psi_p^1(\xi_1))\right) & 0 \\ 0 & Df_2\left(\varphi_{t+q}^2(\xi_2)\right) \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

and recall Diliberto's theorem [5], [6], [8].

THEOREM 3.1 (Diliberto's theorem [5], [8]). *If* $\dot{x} = f(x)$, $x \in \mathbb{R}^2$, $f(\xi) \neq 0$, *and* $t \mapsto x(t, p)$ *is the solution of the differential equation such that* $x(0, p) = p$, *then the homogeneous variational equation*

$$\dot{W} = Df(x(t, \xi))W$$

has a fundamental matrix solution $t \mapsto \Psi(t)$,

$$\Psi(t) = \begin{pmatrix} 1 & \alpha(t, \xi) \\ 0 & \beta(t, \xi) \end{pmatrix}$$

with respect to the moving frame

$$\left\{f(t, \xi),\ ||f(t, \xi)||^{-2} f^{\perp}(t, \xi)\right\},$$

where

$$f(t, \xi) := f(x(t, \xi)),$$

$$\beta(t, \xi) = \exp \int_0^t \operatorname{div} f(s, \xi)\, ds,$$

$$\alpha(t, \xi) = \int_0^t \left\{ \frac{1}{||f||^2} (2\kappa ||f|| - \operatorname{curl} f)\beta \right\} (s, \xi)\, ds,$$

and κ denotes the signed scalar curvature

$$\kappa(t,\xi) := \frac{1}{||f(t,\xi)||^3} f(t,\xi) \wedge Df(t,\xi) f(t,\xi).$$

Also, to compress the notation, we define

$$\begin{aligned}
\alpha_1(s,p) &:= \alpha_1(s,\psi_p^1(\xi_1)), & \beta_1(s,p) &:= \beta_1(s,\psi_p^1(\xi_1)),\\
f_1(s,p) &:= f_1(\varphi_s^1(\psi_p^1(\xi_1))), & f_1^\perp(s,p) &:= f_1^\perp(\varphi_s^1(\psi_p^1(\xi_1))),\\
\alpha_2(s,q) &:= \alpha_2(s,\varphi_q^2(\xi_2)), & \beta_2(s,q) &:= \beta_2(s,\varphi_q^2(\xi_2)),\\
f_2(s,q) &:= f_2(\varphi_{s+q}^2(\xi_2)), & f_2^\perp(s,p) &:= f_2^\perp(\varphi_{s+q}^2(\xi_2)),
\end{aligned}$$

where the subscripts on α and β refer to the functions as defined in Diliberto's theorem for the unperturbed equations $\dot{x}_i = f_i(x_1)$, $i = 1, 2$. Now, the fundamental matrix solution relative to the basis $\mathcal{S}$ for our splitting

$$\begin{aligned}
&\{F(t,p,q), F_1^{\tan}(t,p), F_2^{\tan}(t,q), F^{\nor}(t,q)\}\\
&:= \left\{ \begin{pmatrix} f_1(t,p) \\ f_2(t,q) \end{pmatrix}, \begin{pmatrix} ||f_1(t,p)||^{-2} f_1^\perp(t,p) \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ f_2(t,q) \end{pmatrix}, \begin{pmatrix} 0 \\ ||f_2(t,q)||^{-2} f_2^\perp(t,q) \end{pmatrix} \right\}
\end{aligned}$$

is given by

$$\Phi(t) = \begin{pmatrix} 1 & \alpha_1(t,p) & 0 & 0 \\ 0 & \beta_1(t,p) & 0 & 0 \\ 0 & 0 & 1 & \alpha_2(t,q) \\ 0 & 0 & 0 & \beta_2(t,q) \end{pmatrix}.$$

This means the associated maps a, b, and c defined in §2 reduce as follows:

$a : \mathcal{E}^{\nor}(\psi_p^1(\xi_1), \varphi_q^2(\xi_2)) \to \mathcal{E}^{\tan}(\varphi_t^1(\psi_p^1(\xi_1)), \varphi_t^1(\varphi_q^2(\xi_2)))$ is given by the 2×1 matrix

$$\begin{pmatrix} 0 \\ \alpha_2(t,q) \end{pmatrix};$$

$b : \mathcal{E}^{\nor}(\psi_p^1(\xi_1), \varphi_q^2(\xi_2)) \to \mathcal{E}^{\nor}(\varphi_t^1(\psi_p^1(\xi_1)), \varphi_t^1(\varphi_q^2(\xi_2)))$ is given by the 1×1 matrix $(\beta_2(t,q))$; and

$c : \mathcal{E}^{\tan}(\psi_p^1(\xi_1), \varphi_q^2(\xi_2)) \to \mathcal{E}^{\tan}(\varphi_t^1(\psi_p^1(\xi_1)), \varphi_t^1(\varphi_q^2(\xi_2)))$ is given by the 2×2 matrix

$$\begin{pmatrix} \beta_1(t,p) & 0 \\ 0 & 1 \end{pmatrix}.$$

Step 4 [normal nondegeneracy]. Define the transit time map $T : \mathbb{R}^3 \to \mathbb{R}$ given by $(p,q,\zeta) \mapsto T(p,q,\zeta)$, where $T(p,q,\zeta)$ denotes the time of first return of the point $\Delta(0,p,q,\zeta) \in \Sigma$ to Σ_0, and note $T(p,q,0) \equiv T_{\mathcal{A}}$. To show the normal nondegeneracy, we must show that the kernel of the derivative of the displacement at each point on $\xi \in \Sigma \cap \mathcal{A}$ is two-dimensional. In the present case, since we already know the kernel contains the subspace $\mathcal{E}^{\tan}(\xi)$, it suffices to show that the derivative of the Poincaré map at ξ is not the identity. To prove this we show

$$DP\left(\psi_p^1(\xi_1), \varphi_q^2(\xi_2), 0\right) \begin{pmatrix} 0 \\ f_2^\perp(0,q) \end{pmatrix} \neq \begin{pmatrix} 0 \\ f_2^\perp(0,q) \end{pmatrix}.$$

The vector in the last formula is tangent to the curve

$$\zeta \mapsto \left(\psi_p^1(\xi_1), \psi_\zeta^2(\varphi_q^2(\xi_2))\right)$$

at $\zeta = 0$. So, we must compute the tangent to the curve

$$\zeta \mapsto P\left(\psi_p^1(\xi_1), \psi_\zeta^2(\varphi_q^2(\xi_2))\right) = \left(\varphi^1_{T(p,q,0)}(\psi_p^1(\xi_1)), \varphi^2_{T(p,q,0)}(\psi_\zeta^2(\varphi_q^2(\xi_2)))\right)$$

at $\zeta = 0$. The computation is just an application of Diliberto's theorem. In fact, we obtain

$$DP\left(\psi_p^1(\xi_1), \varphi_q^2(\xi_2), 0\right)\begin{pmatrix} 0 \\ f_2^\perp(0,q) \end{pmatrix} = \begin{pmatrix} 0 \\ D\varphi^2_{T(p,q,0)}(\varphi_q^2(\xi_2)) f_2^\perp(0,q) \end{pmatrix}$$
$$= \begin{pmatrix} 0 \\ ||f_2(0,q)||^2 \left(\alpha_2(T(p,q,0),q) f_2(0,q) + \beta_2(T(p,q,0),q)||f_2(0,q)||^{-2} f_2^\perp(0,q)\right) \end{pmatrix}.$$

The infinitesimal displacement of our vector is given by

$$\mathcal{R}(q) := DP(\psi_p^1(\xi_1), \varphi_q^2(\xi_2), 0)\begin{pmatrix} 0 \\ f_2^\perp(0,0) \end{pmatrix} - \begin{pmatrix} 0 \\ f_2^\perp(0,0) \end{pmatrix}$$
$$= \begin{pmatrix} 0 \\ ||f_2(0,q)||^2 \alpha_2(T_{\mathcal{A}}, q) f_2(0,q) + (\beta_2(T_{\mathcal{A}}, q) - 1) f_2^\perp(0,q) \end{pmatrix}.$$

To see that $\mathcal{R}(q) \not\equiv 0$, we use the following facts: $\beta_2(T_{\mathcal{A}}, q)$ is the characteristic multiplier of Γ and the derivative of the transit time function at Γ is given by $-||f_2(0,0)||\alpha_2(T_{\mathcal{A}}, q)$; see [5] or [6] for more explanation. Since, by the hypotheses stated in §1, either Γ is hyperbolic or Γ belongs to a period annulus such that the derivative of a period function does not vanish at Γ, it follows that $\mathcal{A}$ is normally nondegenerate.

Step 5 [projection to complement of the range of $D\delta(p,q,0,0)$]. It is clear from Step 4 that a two-dimensional complement for the range of $D\delta(p,q,0,0)$, expressed with respect to the basis $\mathcal{S}$ for the splitting over $\mathcal{A}$, is given by

$$\left\{F_1^{\tan}(0,p), \mathcal{R}^\perp(q)\right\},$$

where

$$\mathcal{R}^\perp(q) := \begin{pmatrix} 0 \\ (1 - \beta_2(T(p,q,0),q)) f_2(0,q) + ||f_2(0,q)||^2 \alpha_2(T_{\mathcal{A}}, q) f_2^\perp(0,q) \end{pmatrix}.$$

Moreover, since

$$\left\{F(0,p,q), F_1^{\tan}(0,p), \mathcal{R}(q), \mathcal{R}^\perp(q)\right\}$$

is a basis $\mathcal{T}$ for $\mathbb{R}^4$, the projection from the original splitting to the chosen complement for the range is easy to compute. In fact, there are four functions, each mapping $\mathbb{R}$ to $\mathbb{R}$, given by $q \mapsto k_1(q)$, $q \mapsto k_2(q)$, $q \mapsto B(q)$, and $q \mapsto C(q)$ such that

$$F_2^{\tan}(0,q) = k_1(q)\mathcal{R}(q) + B(q)\mathcal{R}^\perp(q),$$
$$F^{\text{nor}}(0,q) = k_2(q)\mathcal{R}(q) + C(q)\mathcal{R}^\perp(q).$$

Thus, the matrix of the required projection

$$H(p,q) : \left(\mathcal{E} \oplus \mathcal{E}^{\tan} \oplus \mathcal{E}^{\text{nor}}\right)(\Delta(0,p,q,0)) \to \mathbb{R}^2$$

with respect to the (ordered) basis $\mathcal{S}$ on its domain and the (ordered) basis $\mathcal{T}$ on its range is given by the linear map

$$H(p,q)\begin{pmatrix}\varepsilon\\ \tau_1\\ \tau_2\\ \eta\end{pmatrix}=\begin{pmatrix}\tau_1\\ B(q)\tau_2+C(q)\eta\end{pmatrix},$$

where

$$B(q):=\frac{1-\beta_2(T_{\mathcal{A}},q)}{||f_2(0,q)||^4\alpha_2(T_{\mathcal{A}},\xi_2)^2+\left(1-\beta_2(T_{\mathcal{A}},q)\right)^2},$$
$$C(q):=\frac{\alpha_2(T_{\mathcal{A}},q)}{||f_2(0,q)||^4\alpha_2(T_{\mathcal{A}},\xi_2)^2+\left(1-\beta_2(T_{\mathcal{A}},q)\right)^2}.$$

Step 6 [adapted components for perturbation]. The derivative with respect to ϵ at $\epsilon=0$ of the vector field associated with E_ϵ along the unperturbed solution is given by

$$G(t,p,q):=\begin{pmatrix}g_1(t,p,q)\\ g_2(t,p,q)\end{pmatrix}:=\begin{pmatrix}g_1(x_1(t,p),\dot{x}_1(t,p),x_2(t,q),\dot{x}_2(t,q),0)\\ g_2(x_1(t,p),\dot{x}_1(t,p),x_2(t,q),\dot{x}_2(t,q),0)\end{pmatrix},$$

where $t\mapsto(x_1(t,p),x_2(t,q))$ is the unperturbed solution starting at $\Delta(0,p,q,0)$. The vector $G(t,p,q)$ has a unique expression as a linear combination of the vectors in the basis $\mathcal{S}$. In fact, we suppose that

$$G(t,p,q)=\varepsilon F(t,p,q)+\tau_1F_1^{\tan}(t,p)+\tau_2F_2^{\tan}(t,q)+\eta F^{\mathrm{nor}}(t,q)$$

and compute inner products with respect to f_1, $f_1^\perp$, f_2, and $f_2^\perp$ to obtain

$$G^{\tan}(t,p,q):=\begin{pmatrix}\tau_1(t,p,q)\\ \tau_2(t,p,q)\end{pmatrix},$$
$$G^{\mathrm{nor}}(t,p,q):=\eta(t,p,q),$$

where

$$\tau_1(t,p,q)=f_1(t,p)\wedge g_1(t,p,q),$$
$$\tau_2(t,p,q)=\frac{1}{||f_2(t,q)||^2}\langle g_2(t,p,q),f_2(t,q)\rangle-\frac{1}{||f_1(t,p)||^2}\langle g_1(t,p,q),f_1(t,p)\rangle,$$
$$\eta(t,p,q)=f_2(t,q)\wedge g_2(t,p,q).$$

Step 7 [bifurcation function]. Using the definitions of §2 and the results of Steps 3 and 4 we now have

$$\mathcal{M}(p,q)=\int_0^{T_{\mathcal{A}}}b^{-1}(s,q)G^{\mathrm{nor}}(s,p,q)\,ds$$

given by

$$\int_0^{T_{\mathcal{A}}}\frac{1}{\beta_2(s,q)}f_2(t,q)\wedge g_2(t,p,q)\,ds,$$

and

$$\mathcal{N}(p,q) = \int_0^{T_{\mathcal{A}}} c^{-1}(t,p)G^{\tan}(s,p,q) - c^{-1}(t,p)a(t,q)b^{-1}(t,q)G^{\mathrm{nor}}(s,p,q)\,ds$$
$$:= \begin{pmatrix} \mathcal{N}_1(p,q) \\ \mathcal{N}_2(p,q) \end{pmatrix}$$

given by

$$\mathcal{N}_1(p,q) = \int_0^{T_{\mathcal{A}}} \frac{1}{\beta_1(s,p)} f_1(t,p) \wedge g_1(t,p,q)\,ds,$$
$$\mathcal{N}_2(p,q) = \int_0^{T_{\mathcal{A}}} \frac{1}{||f_2(t,q)||^2} \langle g_2(t,p,q), f_2(t,q)\rangle$$
$$- \frac{1}{||f_1(t,p)||^2} \langle g_1(t,p,q), f_1(t,p)\rangle - \frac{\alpha_2(t,q)}{\beta_2(t,q)} f_2(t,q) \wedge g_2(t,p,q)\,ds.$$

Thus, the bifurcation function is given by

$$\mathcal{B}(p,q) = H(p,q) \begin{pmatrix} 0 \\ \mathcal{N}_1(p,q) \\ \mathcal{N}_2(p,q) \\ \mathcal{M}(p,q) \end{pmatrix} = \begin{pmatrix} \mathcal{N}_1(p,q) \\ B(q)\mathcal{N}_2(p,q) + C(q)\mathcal{M}(p,q) \end{pmatrix}.$$

In practice, it is more convenient to clear the nonzero denominator of the second component and use the normalized bifurcation function given by

$$\mathcal{C}(p,q) := \begin{pmatrix} \mathcal{N}_1(p,q) \\ (1-\beta_2(T_{\mathcal{A}},q))\mathcal{N}_2(p,q) + \alpha_2(T_{\mathcal{A}},q)\mathcal{M}(p,q) \end{pmatrix}.$$

Of course, $\mathcal{C}$ and $\mathcal{B}$ have the same set of simple zeros.

4. Applications. We consider an application of our results to the model of a flywheel attached to an elastic support as described in [16]. The model equations are typical for resonance phenomena and are given by

$$\ddot{z} + \omega^2 z = \frac{\epsilon}{m}\left(-f(z) - \beta\dot{z} + q_1\dot{\theta}^2\cos\theta\right) + O(\epsilon^2),$$
$$\ddot{\theta} = \epsilon\left(\frac{1}{J_0}M_1(\dot{\theta}) + q_2\,\mathrm{g}\sin\theta - q_2\omega^2 z\sin\theta\right) + O(\epsilon^2),$$

where z denotes the displacement of the flywheel relative to its support, θ denotes the angular position of the rotating flywheel relative to the (upward) vertical, g is the gravitational constant, and M_1 is the motor characteristic. The remaining parameters are all constant with, of course, ϵ being a small parameter. To apply the results of §3, we write the model equation as a first-order system using the transformation $x = \dot{\theta}\cos\theta$, $y = \dot{\theta}\sin\theta$, and assuming $\dot{\theta} > 0$ to obtain

$$\dot{z} = -\omega w,$$
$$\dot{w} = \omega z - \epsilon g(z,w,x,y),$$
$$\dot{x} = -y\sqrt{x^2+y^2} + \epsilon\frac{x}{\sqrt{x^2+y^2}}h(z,w,x,y) + O(\epsilon^2),$$
$$\dot{y} = x\sqrt{x^2+y^2} + \epsilon\frac{y}{\sqrt{x^2+y^2}}h(z,w,x,y) + O(\epsilon^2),$$

where

$$g(z,w,x,y) = \frac{1}{m\omega}\left(-f(z) - \beta w + q_1 x\sqrt{x^2+y^2}\right),$$
$$h(z,w,x,y) = \frac{1}{J_0}M_1\left(\sqrt{x^2+y^2}\right) + q_2\,\mathrm{g}\frac{y}{\sqrt{x^2+y^2}} - q_2\omega^2\frac{yz}{\sqrt{x^2+y^2}}.$$

The transformation to (x,y) variables is geometrically a coordinate chart on the tangent bundle of the circle

$$\{(\theta,\dot\theta) \mid \theta \in \mathbb{S}^1 \text{ and } \dot\theta \in \mathbb{R}\}.$$

The chart does not contain the zero section ($\dot\theta = 0$), but this set is not near the resonance. In fact, the first oscillator is linear with its period annulus A having period $2\pi/\omega$, while the second oscillator has a period annulus at the origin whose period function is given by $r \mapsto 2\pi/r$, where $r := \sqrt{x^2+y^2}$. The primary resonance is given by $r = \omega$. In other words, the resonant periodic solution Γ in the second oscillator lies on the invariant circle of radius ω. With

$$f_1(z,w) := -w\frac{\partial}{\partial z} + z\frac{\partial}{\partial w}, \qquad f_2(x,y) := -y\sqrt{x^2+y^2}\frac{\partial}{\partial x} + x\sqrt{x^2+y^2}\frac{\partial}{\partial y},$$

$\xi_1 := (1,0)$, and $\xi_2 := (\omega,0)$, we find that the solution of the unperturbed system with initial value $(\phi_p^1(\xi_1), \varphi_q^2(\xi_2))$ is given by

$$\begin{aligned} z(t,p) &= \mathrm{e}^{-p}\cos(\omega t), & w(t,p) &= \mathrm{e}^{-p}\sin(\omega t),\\ x(t,q) &= \omega\cos(\omega(t+q)), & y(t,q) &= \omega\sin(\omega(t+q)). \end{aligned}$$

From the results of §3, the bifurcation function is

$$\begin{aligned} \mathcal{B}(p,q) &= \left(\omega\int_0^{2\pi/\omega} wg(z,w,x,y)\,dt,\ -\int_0^{2\pi/\omega}(x^2+y^2)h(z,w,x,y)\,dt\right)\\ &= \left(\frac{\pi\omega}{m}\,\mathrm{e}^{-2p}(\beta - q_1\omega\,\mathrm{e}^{p}\sin(\omega q)),\ -\frac{\pi\omega^2}{J_0}(2M_1(1/\omega) - J_0q_2\omega^2\,\mathrm{e}^{-p}\sin(\omega q))\right). \end{aligned}$$

The bifurcation function has either 0, 1, or 2 zeros depending on the values of the parameters. The zeros are obtained when the following two equations can be solved for both p and q :

$$\mathrm{e}^{2p} = \frac{\beta\omega J_0 q_2}{2M_1(1/\omega)q_1}, \qquad \sin\omega q = \frac{\beta}{\omega q_1}\sqrt{\frac{2M_1(1/\omega)q_1}{\beta\omega J_0 q_2}}.$$

If we choose $\beta = \omega = J_0 = q_1 = q_2 = 1$ and $M_1(1) = 1/4$, then these equations reduce to

$$\mathrm{e}^{p} = \sqrt{2}, \qquad \sin q = \frac{1}{\sqrt{2}}.$$

In this case $p = \ln\sqrt{2}$, $q = \pi/4$, $3\pi/4$ are zeros of the bifurcation function. Since the bifurcation function can be normalized to

$$(p,q) \mapsto \left(\beta - q_1\omega\,\mathrm{e}^{p}\sin(\omega q),\ 2M_1(1/\omega) - J_0q_2\omega^2\,\mathrm{e}^{-p}\sin(\omega q)\right),$$

a map whose Jacobian is

$$2q_1q_2J_0\omega^2\sin(\omega q)\cos(\omega q)$$

has the zeros of the bifurcation function that are simple except when $\sin(\omega q) = \pm 1$. In particular, the zeros of the numerical example are simple and, by the results of §3, there are two bifurcating families of periodic solutions in the model equations for the flywheel with elastic support. We emphasize that although the analysis uses only the $O(\epsilon)$ terms of the model, our result is valid for small ϵ for the full model equations; compare to [16].

The analysis just given is prototypical. However, there are other resonances to consider. Using the notation defined above, the general resonance relation is given by

$$K_1\frac{2\pi}{\omega} = K_2\frac{2\pi}{r}$$

or $r = K_2\omega/K_1$. On the resonant orbit,

$$x(t,q) = \frac{K_2}{K_1}\omega\cos\frac{K_2}{K_1}\omega(t+q), \qquad y(t,q) = \frac{K_2}{K_1}\omega\sin\frac{K_2}{K_1}\omega(t+q).$$

Thus, the first component of the bifurcation function is given by

$$\begin{aligned}
&\omega\int_0^{K_1 2\pi/\omega} wg(z,w,x,y)\,dt \\
&= \frac{1}{m}\int_0^{K_1 2\pi/\omega} \mathrm{e}^{-p}\sin\omega t\left(-f(z) - \beta\,\mathrm{e}^{-p}\sin\omega t + q_1\left(\frac{K_2}{K_1}\omega\right)^2\cos\frac{K_2}{K_1}\omega(t+q)\right)dt \\
&= \frac{1}{m}\,\mathrm{e}^{-2p}\frac{K_1\pi}{\omega} + \frac{q_1}{m}\left(\frac{K_2}{K_1}\omega\right)^2\mathrm{e}^{-p}I_0(q),
\end{aligned}$$

where

$$I_0(q) := \int_0^{K_1 2\pi/\omega}\sin\omega t\cos\frac{K_2}{K_1}\omega(t+q)\,dt.$$

As $I_0(q)$ is nonzero only when $K_1 = K_2$, nondegenerate bifurcation to periodic orbits occurs only for the primary resonance.

Up to this point we have assumed that several forces are small. To illustrate the possibility of relaxing this hypothesis, consider the rotor to be influenced strongly by a "gravitational" force. It is convenient to measure the inclination of the rotor by the angle of displacement from the direction of the gravitational force, downward vertical, i.e., we use the angle $\psi = -\theta - \pi$. The model equations (up to first order in ϵ) become (to first order)

$$\begin{aligned}
\ddot{z} + \omega^2 z &= \frac{\epsilon}{m}\left(-f(z) - \beta\dot{z} + q_1\dot{\psi}^2(-\cos\psi)\right), \\
\ddot{\psi} &= -\epsilon\left(\frac{1}{J_0}M_1(-\dot{\psi}) + q_2\,\mathrm{g}\sin\psi - q_2\omega^2 z\sin\psi\right).
\end{aligned}$$

To study the strong gravitational effect we assume $\mathrm{g} := G/\epsilon$ and transform the independent variable by $\tau = t\sqrt{q_2G}$ to obtain

$$\begin{aligned}
q_2Gz'' + \omega^2 z &= -\frac{\epsilon}{m}\left(f(z) + \beta z'\sqrt{q_2G} + q_1q_2G(\psi')^2\cos\psi\right), \\
q_2G\psi'' + q_2G\sin\psi &= -\epsilon\left(\frac{1}{J_0}M_1(-\psi'\sqrt{q_2G}) - q_2\omega^2 z\sin\psi\right),
\end{aligned}$$

which we rewrite in the form

$$\ddot{z} + \Omega^2 z = \epsilon g(z, \dot{z}, \theta, \dot{\theta}),$$
$$\ddot{\theta} + \sin\theta = \epsilon h(z, \dot{z}, \theta, \dot{\theta}),$$

where

$$g(z, \dot{z}, \theta, \dot{\theta}) := -\left(F(z) + \lambda\dot{z} + A\dot{\theta}^2 \cos\theta\right),$$
$$h(z, \dot{z}, \theta, \dot{\theta}) := -\left(M(\dot{\theta}) - Bz\sin\theta\right),$$

and the new parameters and functions have obvious meaning. In particular,

$$\lambda := \frac{\beta}{m}(q_2 G)^{-1/2}, \qquad \Omega := \omega(q_2 G)^{-1/2}, \qquad A := \frac{q_1}{m}, \qquad B := \frac{\omega^2}{G}.$$

The first oscillator corresponding to the elastic support has the entire punctured phase plane as an isochronous period annulus with period $2\pi/\Omega$. In fact, if we view the system in the phase plane as

$$\dot{z} = -\Omega w, \qquad \dot{w} = \Omega z - \frac{\epsilon}{\Omega} g(z, \dot{z}, \theta, \dot{\theta}),$$

then the solution of the unperturbed oscillator with initial value $(e^{-p}, 0)$ is given by

$$z(t, p) = e^{-p}\cos\Omega t, \qquad w(t, p) = e^{-p}\sin\Omega t.$$

The second oscillator models the rotor influenced by a gravitational field. The unperturbed second oscillator is a mathematical pendulum. It has a period annulus (with strictly monotone period function) surrounding the origin of the phase plane. This period annulus corresponds to the nonrotational oscillations of the pendulum. Also, there is a period annulus in the phase cylinder (with strictly monotone period function) corresponding to the rotational oscillations. Thus we have the hypotheses required to apply the theoretical results of §3. The analysis to follow uses elliptic functions. Perhaps this can be avoided.

To compute the bifurcation function we require the time-dependent solutions of the mathematical pendulum given in the phase plane by the first-order system

$$\dot{\theta} = v, \qquad \dot{v} = -\sin\theta.$$

For the convenience of the reader and to fix notation, we will outline the usual derivation.

Consider the period annulus in the phase plane. The mathematical pendulum has the first integral $I := v^2/2 - \cos\theta$. For a periodic trajectory Γ, let $(a, 0)$ denote the coordinates of its intersection with the θ-axis. On Γ the energy is $I \equiv -\cos a$ and $\dot{\theta}^2 = 2(\cos\theta - \cos a)$. By integration and the change of variables $\sin(\theta/2) = \sin(a/2)\sin\varphi$, we find

$$t = \int_0^{\varphi(t)} \frac{1}{\sqrt{1 - \sin^2(a/2)\sin^2 s}}\, ds,$$

where $\theta(t)$ is the solution of the mathematical pendulum with the initial value

$$(\theta(0), \dot{\theta}(0)) = (0, 2\sin(a/2)).$$

Or, in terms of Jacobian elliptic functions (cf. [4], [20]) where the elliptic modulus is $k := \sin(a/2)$, we find

$$\sin\varphi(t) = \operatorname{sn}[t, k]$$

and, using the trigonometric double angle formulas,

$$\cos\theta(t) = 1 - 2k^2 \operatorname{sn}^2[t, k].$$

Also, the period of Γ is given by

$$4\int_0^{\pi/2} \frac{1}{\sqrt{1 - k^2 \sin^2 s}}\, ds = 4K(k),$$

where $K(k)$ is the complete elliptic integral of the first kind. Since $t \mapsto \operatorname{sn}(t)$ has real period $4K$ (here and hereafter if the elliptic modulus is not given explicitly it is understood to be $k = \sin(a/2)$), the periodic orbit Γ is resonant when there are relatively prime positive integers K_1 and K_2 such that

$$K_1 \frac{2\pi}{\Omega} = K_2 4K(k).$$

Under this assumption and in view of the first-order system

$$\begin{aligned} \dot{z} &= -\Omega w, \\ \dot{w} &= \Omega z - \frac{\epsilon}{\Omega} g, \\ \dot{\theta} &= v, \\ \dot{v} &= -\sin\theta + \epsilon h, \end{aligned}$$

the bifurcation function for a nonrotational resonance is given by

$$\mathcal{B}(p, q) = \left(\int_0^{K_1 2\pi/\Omega} wg\, dt,\ \int_0^{K_1 2\pi/\Omega} vh\, dt \right),$$

where q is the coordinate on Γ introduced by using the solution $t \mapsto \theta(t + q)$ for $0 \le q < 4K(k)$. The components of the bifurcation function are computed as follows:

$$\begin{aligned} \int_0^{K_1 2\pi/\Omega} wg\, dt &= K_1 \pi\lambda\, \mathrm{e}^{-2p} - A\, \mathrm{e}^{-p} I_1(q), \\ \int_0^{K_1 2\pi/\Omega} vh\, dt &= B\, \mathrm{e}^{-p} I_2(q) - I_3(q), \end{aligned}$$

where

$$\begin{aligned} I_1(q) &:= \int_0^{K_1 2\pi/\Omega} (\dot{\theta}(t + q))^2 \cos\theta(t + q) \sin\Omega t\, dt, \\ I_2(q) &:= \int_0^{K_1 2\pi/\Omega} \dot{\theta}(t + q) \sin\theta(t + q) \cos\Omega t\, dt, \\ I_3(q) &:= \int_0^{K_1 2\pi/\Omega} \dot{\theta}(t + q) M(\dot{\theta}(t + q))\, dt. \end{aligned}$$

The integral I_3 depends on the static characteristic of the motor and the damping associated with the rotational motion as encoded in the function M. As a typical example and for definiteness in the computation, we take M to be linear,

$$M(\dot{\theta}) := m_1 + m_2\dot{\theta};$$

more general model functions can be handled in a similar manner. For the linear case, we have the following proposition:

$$\begin{aligned} I_3(q) &= \int_0^{K_1 2\pi/\Omega} m_1\dot{\theta} + m_2\dot{\theta}^2\, dt \\ &= 2m_2 \int_0^{K_1 2\pi/\Omega} \cos\theta(t+q) - \cos a\, dt \\ &= -4m_2K_1\frac{\pi}{\Omega}\cos a + 2m_2\int_0^{K_2 4K} \cos\theta(t)\, dt \\ &= -2m_2K_2 4K\cos a + 2m_2\int_0^{K_2 4K} 1 - 2k^2\,\mathrm{sn}^2(t)\, dt. \end{aligned}$$

Formula 310.02 in [4] can be used to evaluate the integral with integrand $\mathrm{sn}^2(t)$ to obtain

$$\begin{aligned} I_3(q) &= -8m_2K_2(1+\cos a)K(k) + 4m_2E(\mathrm{am}\,[K_2 4K(k), k], k) \\ &= -16m_2K_2(1-k^2)K(k) + 4m_2E(\mathrm{am}\,[K_2 4K(k), k], k), \end{aligned}$$

where $E(\varphi, k)$ denotes the normal elliptic integral of the second kind and $\mathrm{am}\,[u, k]$ is the amplitude; see [4]. Using [4, formulas 113.02 and 122.06], we obtain

$$E(\mathrm{am}\,[K_2 4K(k), k], k) = 4K_2E(k),$$

where $E(k)$ is the complete elliptic integral of the second kind. Thus

$$I_3(q) = 16m_2K_2(E(k) - (1-k^2)K(k)).$$

Note that for the linear static motor characteristic, $q \mapsto I_3(q)$ is constant. Moreover,

$$I_3^* := \frac{1}{m_2}I_3(q) = \int_0^{K_1 2\pi/\Omega} \dot{\theta}^2\, dt > 0.$$

For the integrals I_1 and I_2 we have the following identity.

IDENTITY 4.1.

$$\frac{3}{2}\Omega I_1(q) = (\cos a - \Omega^2)I_2(q).$$

Proof. Define $\eta := K_1 2\pi/\Omega$ and compute the following:

$$\begin{aligned} I_1(q) &= \int_0^{\eta} 2(\cos\theta - \cos a)\cos\theta\sin\Omega t\, dt \\ &= 2\int_0^{\eta} \cos^2\theta\sin\Omega t\, dt - 2\cos a\int_0^{\eta} \cos a\sin\Omega t\, dt, \end{aligned}$$

$$
\begin{aligned}
I_2(q) &= -\frac{1}{\Omega}\int_0^\eta (\ddot{\theta}\sin\theta + \dot{\theta}^2\cos\theta)\sin\Omega t\,dt \\
&= -\frac{1}{\Omega}\int_0^\eta \ddot{\theta}\sin\theta\sin\Omega t\,dt - \frac{1}{\Omega}I_1(q) \\
&= \frac{1}{\Omega}\int_0^\eta \sin^2\theta\sin\Omega t\,dt - \frac{1}{\Omega}I_1(q) \\
&= -\frac{1}{\Omega}\int_0^\eta \cos^2\theta\sin\Omega t\,dt - \frac{1}{\Omega}I_1(q) \\
&= -\frac{1}{\Omega}\left(\frac{1}{2}I_1(q) + \cos a\int_0^\eta \cos\theta\sin\Omega t\,dt\right) - \frac{1}{\Omega}I_1(q) \\
&= -\frac{3}{2\Omega}I_1(q) - \frac{\cos a}{\Omega}\left(-\int_0^\eta (-\dot{\theta}\sin\theta)\left(-\frac{1}{\Omega}\cos\Omega t\right)\,dt\right) \\
&= -\frac{3}{2\Omega}I_1(q) + \frac{\cos a}{\Omega^2}I_2(q). \qquad \square
\end{aligned}
$$

Also, with the definition

$$I_c := \int_0^{K_1 2\pi/\Omega} \cos\theta(t)\cos\Omega t\,dt$$

we have a second identity.

IDENTITY 4.2.

$$I_2(q) = \Omega I_c \sin\Omega q.$$

Proof. Define $\eta := K_1 2\pi/\Omega$ and compute the following:

$$
\begin{aligned}
I_2(q) &= -\int_0^\eta \frac{d}{dt}(\cos\theta(t+q))\cos\Omega t\,dt \\
&= -\Omega\int_0^\eta \cos\theta(t+q)\sin\Omega t\,dt \\
&= -\Omega\int_0^\eta \cos\theta(t)\sin\Omega(t-q)\,dt \\
&= -\Omega\cos\Omega q\int_0^\eta \cos\theta(t)\sin\Omega t\,dt + \Omega\sin\Omega q\int_0^\eta \cos\theta(t)\cos\Omega t\,dt.
\end{aligned}
$$

Since $t \mapsto \cos\theta(t)$ is an even function,

$$\int_0^\eta \cos\theta(t)\sin\Omega t\,dt = 0. \qquad \square$$

Using the identities just obtained, we have

$$\mathcal{B}(p,q) = \left(K_1\pi\lambda\,\mathrm{e}^{-2p} - \frac{2}{3}AI_c\,\mathrm{e}^{-p}(\cos a - \Omega^2)\sin\Omega q,\ -I_3 + BI_c\Omega\,\mathrm{e}^{-p}\sin\Omega q\right).$$

Thus, (p,q) is a zero of the bifurcation function if and only if this ordered pair is a solution of the bifurcation equations

$$
\begin{aligned}
K_1\pi\lambda - \frac{2}{3}AI_c(\cos a - \Omega^2)\,\mathrm{e}^p\sin\Omega q &= 0, \\
I_3 - BI_c\Omega e^{-p}\sin\Omega q &= 0.
\end{aligned}
$$

Such a zero is simple provided that

$$\frac{4}{3}ABI_c^2(\cos a - \Omega^2)\sin\Omega q\cos\Omega q \neq 0.$$

To show that the bifurcation is nondegenerate, we must show that $I_c \neq 0$. It turns out that the validity of this condition depends on the resonance. This is the content of the following proposition.

PROPOSITION 4.3. *If K_1 and K_2 are relatively prime positive integers such that $K_1 2\pi/\Omega = K_2 4K(k)$, then for I_c to be nonvanishing it is necessary and sufficient that $K_2 = 1$ and $K_1 = 2n$ for some positive integer n. In case this condition holds,*

$$I_c = 4\left(\frac{\pi^2 K_1}{K(k)}\right)\frac{\mathbf{q}^{K_1/2}}{1-\mathbf{q}^{K_1}} = 8\pi K_2\Omega\frac{\mathbf{q}^{K_1/2}}{(1-\mathbf{q}^{K_1})},$$

where $\mathbf{q} := \mathrm{e}^{-\pi K'/K}$ is Jacobi's nome, [4, p. 315].

Proof. The proposition follows from the Fourier series representation of $u \mapsto \mathrm{sn}^2(u)$ given by

$$(kK)^2\,\mathrm{sn}^2(u) = K^2 - KE - 2\pi^2\sum_{n=1}^{\infty}\frac{n\mathbf{q}^n}{1-\mathbf{q}^{2n}}\cos 2nx,$$

where $x := \pi u/(2K)$. (This formula is stated without proof in [20, p. 520]. A second Fourier series expansion in [4, formula 911.01] seems to be incorrect. Thus, even though a reference for the formula exists, we will verify this series representation below.) Define $\eta := K_1 2\pi/\Omega$. To prove the proposition, compute

$$\begin{aligned} I_c &= 4\left(\frac{\pi}{K}\right)^2\sum_{n=1}^{\infty}\frac{n\mathbf{q}^n}{1-\mathbf{q}^{2n}}\int_0^{\eta}\cos\left(\frac{n\pi}{K}u\right)\cos\Omega u\,du \\ &= 4\left(\frac{\pi}{K}\right)^2\sum_{n=1}^{\infty}\frac{n\mathbf{q}^n}{1-\mathbf{q}^{2n}}\int_0^{\eta}\cos\left(2n\frac{K_2}{K_1}\Omega u\right)\cos\Omega u\,du. \end{aligned}$$

After the substitution $v := \Omega u/K_1$, we obtain

$$I_c = 4\left(\frac{\pi}{K}\right)^2\sum_{n=1}^{\infty}\frac{n\mathbf{q}^n}{1-\mathbf{q}^{2n}}\frac{K_1}{\Omega}\int_0^{2\pi}\cos 2nK_2 v\cos K_1 v\,dv.$$

Thus, I_c vanishes unless $2nK_2 = K_1$. In particular, K_1 must be even and K_2 must be a factor of K_1. Since K_1 and K_2 are relatively prime, $K_2 = 1$. If $I_c \neq 0$, then

$$\begin{aligned} I_c &= 4\left(\frac{\pi}{K}\right)^2\left(\frac{K_1}{\Omega}\right)\left(\frac{K_1}{2}\right)\frac{\mathbf{q}^{K_1/2}}{1-\mathbf{q}^{K_1}}\pi \\ &= 4\frac{\pi^2 K_1}{K}\frac{\mathbf{q}^{K_1/2}}{1-\mathbf{q}^{K_1}} \end{aligned}$$

as required.

To verify the Fourier series expansion we compute the value of

$$J := \int_{-\pi}^{\pi}\mathrm{sn}^2\left(\frac{2K}{\pi}x\right)\mathrm{e}^{imx}\,dx, \qquad m \neq 0$$

by contour integration around the parallelogram in the complex plane with vertices $-\pi$, π, $\pi\tau$, and $\pi\tau - 2\pi$, where $\tau := iK'/K$; cf. [20, p. 510]. Using the fact that $u \mapsto \operatorname{sn}(u)$ is doubly periodic with periods $4K$ and $2iK'$, and $x \mapsto \mathrm{e}^{imx}$ is periodic with period 2π, the path integrals along the edges of the parallelogram given by $[\pi, \pi\tau]$ and $[\pi\tau - 2\pi, -\pi]$ cancel. Also, an easy computation shows that the integral along the edge $[\pi\tau, \pi\tau - 2\pi]$ is $-\mathrm{e}^{im\pi\tau}\,\mathrm{e}^{im\pi}J$. Thus,

$$\left(1 - \mathrm{e}^{im\pi\tau}\,\mathrm{e}^{im\pi}\right) J = 2\pi i \sum(\text{residues}).$$

The poles of $u \mapsto \operatorname{sn}(u)$ reside at the points in the complex plane congruent to iK' and $2K + iK'$ modulo the periods of sn. It follows that $\operatorname{sn}^2(2Kx/\pi)\,\mathrm{e}^{imx}$ has exactly two poles in the parallelogram. These poles are at the points $\pi\tau/2$ and $\pi\tau/2 - \pi$. To compute the residues, start with the Maclaurin series for $u \mapsto \operatorname{sn}(u)$ given by

$$\operatorname{sn}(u) = u + O(u^3)$$

and the identity

$$\operatorname{sn}(u + iK') = \frac{1}{k\operatorname{sn}(u)}$$

to obtain

$$\operatorname{sn}(u + iK') = \frac{1}{ku} + O(u).$$

Set $u + iK' = 2Kx/\pi$ to get

$$\operatorname{sn}\left(\frac{2K}{\pi}x\right) = \frac{\pi}{2kK(x - \pi\tau/2)} + O(x - \pi\tau/2)$$

and compute

$$\operatorname{sn}^2\left(\frac{2K}{\pi}x\right)\mathrm{e}^{imx} = \left(\frac{\pi}{2kK}\right)^2 \frac{\mathrm{e}^{im\pi\tau/2}}{(x - \pi\tau/2)^2} + \left(\frac{\pi}{2kK}\right)^2 \frac{im\,\mathrm{e}^{im\pi\tau/2}}{(x - \pi\tau/2)} + O(1).$$

Thus, the residue at $\pi\tau/2$ is

$$\left(\frac{\pi}{2kK}\right)^2 im\,\mathrm{e}^{im\pi\tau/2} = \left(\frac{\pi}{2kK}\right)^2 im\mathbf{q}^{m/2}.$$

Use the identity

$$\operatorname{sn}(u - 2K + iK') = -\operatorname{sn}(u + iK')$$

and a similar computation to compute the residue at $-\pi + \pi\tau/2$. We find that this residue is

$$\left(\frac{\pi}{2kK}\right)^2 im\,\mathrm{e}^{-im\pi}\mathbf{q}^{m/2}.$$

From this it follows that

$$J = -\frac{\pi^3}{2(kK)^2}\frac{m\mathbf{q}^{m/2}}{1 - \mathbf{q}^m\,\mathrm{e}^{im\pi}}\left(1 + \mathrm{e}^{-im\pi}\right).$$

Thus, the Fourier coefficient corresponding to J vanishes unless $m = 2n$, in which case

$$J = -2\frac{\pi^3}{(kK)^2}\frac{n\mathbf{q}^n}{1-\mathbf{q}^{2n}}.$$

Since $x \mapsto \mathrm{sn}^2(2Kx/\pi)$ is even, its Fourier series is a cosine series. In fact, the Fourier coefficient of $\cos 2nx$ is the real part of J/π. Since J is real,

$$\mathrm{sn}^2\left(\frac{2K}{\pi}x\right) = a_0 - 2\left(\frac{\pi}{kK}\right)^2 \sum_{n=1}^{\infty} \frac{n\mathbf{q}^n}{1-\mathbf{q}^{2n}} \cos 2nx.$$

The constant term a_0 can be shown to agree with the stated formula, but, since we do not require its value here, the proof is left to the reader. □

By the proposition we see that there are (under appropriate choices of the constant parameters) bifurcating families of periodic solutions for the full model equations at each nonrotational periodic motion of the gravitationally influenced rotor, whose period is an even multiple of the natural period of the support oscillator. In fact, if we impose the nondegeneracy conditions $K_1 = 2n$ and $K_2 = 1$, then, by eliminating $\sin \Omega q$ from the bifurcation equations, we find

$$\mathrm{e}^{2p} = \left(\frac{3\pi\lambda B\Omega}{A}\right)\frac{n}{m_2 I_3^*(\cos a - \Omega^2)}.$$

Thus, we can solve for p provided $m_2(\cos a - \Omega^2) > 0$. Assuming this condition is satisfied and inserting e^p into the second bifurcation equation, we find that there are two solutions for q provided $-1 < \Delta < 1$ for

$$\Delta := \left(\frac{3n\pi\lambda}{AB\Omega}\right)^{1/2}\left(\frac{I_3^*}{m_2(\cos a - \Omega^2)}\right)^{1/2}\left(\frac{m_2}{I_c}\right).$$

The question arises as to how many resonant periodic solutions of the unperturbed mathematical pendulum correspond to nondegenerate bifurcation points. It is clearly possible to obtain any preassigned finite number of simultaneous bifurcations. However, it is not possible to have infinitely many. To have infinitely many bifurcation points for a fixed set of parameter values it is necessary that Δ remain bounded in the unit interval for infinitely many integers n such that $n = \Omega K(k)/\pi$. To show that this is not the case, note that $k \to 1$ as $n \to \infty$, and use the computations made above for I_3^* and I_c, together with the fact that $\cos a = 1 - 2k^2$, to compute

$$\Delta = -\sqrt{n}\, m_2 k^2 \left(\frac{3\lambda}{16AB\Omega^3\pi}\right)^{1/2}\left(\frac{1-\mathbf{q}^{2n}}{\mathbf{q}^n}\right)\left(\frac{I_3^*(k)}{m_2(1-2k^2-\Omega^2)}\right)^{1/2}.$$

We claim that Δ grows without bound as $k \to 1$. First note that as $k \to 1$,

$$\mathbf{q}^n = \mathrm{e}^{-\Omega K(\sqrt{1-k^2})} \to \mathrm{e}^{-\Omega\pi/2},$$

so the term $(1 - \mathbf{q}^{2n})/(\mathbf{q}^n)$ remains bounded. Also, as $k \to 1$ we have $K(k) - \ln(4/\sqrt{1-k^2}) \to 0$. Using these facts and the expression for $I_3^*(k)$, it follows that $I_3^*(k)$ remains bounded as $k \to 1$. Thus, all terms except the $\sqrt{n}$ term remain bounded. It follows that $\Delta \to \infty$ as $k \to 1$ and $n \to \infty$. However, the fact that infinitely

many different resonances can lead to nondegenerate first-order bifurcation to periodic solutions is in marked contrast to the case when the gravitational forces are considered small and the only nondegenerate bifurcation occurs for the primary resonance.

To analyze the rotational motion of the rotor, recall that the mathematical pendulum system defined on the phase plane is given by

$$\dot{\theta} = v, \qquad \dot{v} = -\sin\theta + \epsilon h.$$

The rotational motions are naturally defined on the phase cylinder that is obtained from the phase plane by viewing the variable θ modulo 2π. There are two families of periodic solutions corresponding to $\dot{\theta} < 0$ and $\dot{\theta} > 0$. For definiteness we will treat the case $\dot{\theta} < 0$; the other case is similar. In particular, since we have changed the coordinates of the model equation by $\theta \to -\theta - \pi$, a positive rotation in the original model equations corresponds to a negative rotation here. It is convenient to choose the (symplectic) coordinate chart on the phase cylinder given by the transformations $x = \sqrt{-v}\cos\theta$, $y = \sqrt{-v}\sin\theta$. The chart for the second case would be $x = \sqrt{v}\sin\theta$, $y = \sqrt{v}\cos\theta$. This choice of coordinates ensures that the divergence of the transformed vector field vanishes and the function $\beta_2(t,p)$ defined in §3 is zero. In the (x,y) plane, the phase plane system becomes

$$\dot{x} = y(x^2+y^2) + \frac{1}{2}xy(x^2+y^2)^{-3/2} - \frac{\epsilon}{2}\frac{x}{x^2+y^2}h,$$
$$\dot{y} = -x(x^2+y^2) + \frac{1}{2}y^2(x^2+y^2)^{-3/2} - \frac{\epsilon}{2}\frac{y}{x^2+y^2}h,$$

where

$$h = -\left(M(\dot{\theta}) - Bz\sin\theta\right) = -\left(M(-(x^2+y^2)) - B\frac{yz}{\sqrt{x^2+y^2}}\right).$$

We study the above system coupled as before to the support oscillator given by

$$\dot{z} = -\Omega w, \qquad \dot{w} = \Omega z - \frac{\epsilon}{\Omega}g,$$

where

$$g = -\left(F(z) + \lambda\dot{z} + A\dot{\theta}^2\cos\theta\right)$$
$$= -\left(F(z) - \lambda\Omega w + Ax(x^2+y^2)^{3/2}\right).$$

Since the rotational motions correspond to curves in the phase plane which do not intersect the θ axis, it is convenient to consider the v axis as a section for the flow. On the trajectory passing through the point in the phase plane with coordinates $(0,b)$, $|b| > 2$, the first integral $I := v^2/2 - \cos\theta$ has the constant value $I \equiv b^2/2 - 1$. The case $\dot{\theta} < 0$ corresponds to $b < 2$ and we have

$$\frac{1}{2}\dot{\theta}^2 = \cos\theta + \frac{1}{2}b^2 - 1.$$

Define $\varphi := \theta/2$ and $k := 2/|b|$ to obtain equivalently

$$\dot{\varphi}^2 = \frac{1}{k^2}\left(1 - k^2\sin^2\varphi\right),$$

so that

$$\int_0^{\varphi(t)} \frac{1}{\sqrt{1-k^2\sin^2 s}}\,ds = \operatorname{sgn}(\mathrm{b})\frac{t}{k}.$$

In terms of Jacobian elliptic functions, we have

$$\begin{aligned} \varphi(t) &= \operatorname{sgn}(\mathrm{b})\operatorname{am}[t/k,k], & \dot\varphi(t) &= \operatorname{sgn}(\mathrm{b})\operatorname{dn}[t/k,k]/k,\\ \cos\varphi(t) &= \operatorname{cn}[t/k,k], & \sin\varphi(t) &= \operatorname{sgn}(\mathrm{b})\operatorname{sn}[t/k,k], \end{aligned}$$

or, using the trigonometric double angle formulas,

$$\begin{aligned} \theta(t) &= 2\operatorname{sgn}(\mathrm{b})\,\operatorname{am}[t/k,k], & \dot\theta(t) &= 2\operatorname{sgn}(\mathrm{b})\,\operatorname{dn}[t/k,k]/k,\\ \cos\theta(t) &= 1-2\operatorname{sn}^2[t/k,k], & \sin\theta(t) &= 2\operatorname{sgn}(\mathrm{b})\,\operatorname{sn}[t/k,k]\,\operatorname{cn}[t/k,k]. \end{aligned}$$

Using these formulas, the definition of the phase cylinder, and $b<0$, we have

$$\begin{aligned} x(t) &= \left(\frac{2}{k}\operatorname{dn}[t/k,k]\right)^{1/2}\left(1-2\operatorname{sn}^2[t/k,k]\right),\\ y(t) &= \left(\frac{2}{k}\operatorname{dn}[t/k,k]\right)^{1/2} 2\operatorname{sn}[t/k,k]\operatorname{cn}[t/k,k]. \end{aligned}$$

Also, observe that the period T of the periodic solution on the phase cylinder with initial value $(0,b)$ is given by $\theta(T/2)=\operatorname{sgn}(\mathrm{b})\pi$. Thus,

$$T = 2k\operatorname{am}^{-1}[\pi/2,k] = 2kK(k)$$

and the resonance relation is given by

$$K_1\frac{2\pi}{\Omega} = K_2 2kK(k).$$

Using the results of §3,

$$\mathcal{B} = \left(\int_0^{K_1 2\pi/\Omega} wg\,dt,\ -\frac{1}{2}\int_0^{K_1 2\pi/\Omega}(x^2+y^2)h\,dt\right).$$

Initially, it is preferable to express the components of $\mathcal{B}$ in phase plane coordinates:

$$\begin{aligned} \int_0^{K_1 2\pi/\Omega} wg\,dt &= K_1\pi\lambda\,\mathrm{e}^{-2p} - A\,\mathrm{e}^{-p}I_1^r(q),\\ -\frac{1}{2}\int_0^{K_1 2\pi/\Omega}(x^2+y^2)h\,dt &= -\frac{1}{2}I_3^r(q) + \frac{1}{2}B\,\mathrm{e}^{-p}I_2^r(q), \end{aligned}$$

where

$$\begin{aligned} I_1^r(q) &:= \int_0^{K_1 2\pi/\Omega}(\dot\theta(t+q))^2\cos\theta(t+q)\sin\Omega t\,dt,\\ I_2^r(q) &:= \int_0^{K_1 2\pi/\Omega}\dot\theta(t+q)\sin\theta(t+q)\cos\Omega t\,dt,\\ I_3^r(q) &:= \int_0^{K_1 2\pi/\Omega}\dot\theta(t+q)M(\dot\theta(t+q))\,dt. \end{aligned}$$

Also, we define

$$I_c^r(q) := \int_0^{K_1 2\pi/\Omega} \cos\theta(t)\cos\Omega t\,dt.$$

As in the case of the nonrotational motions, we have the following identities.

IDENTITY 4.4.

$$\frac{3}{2}\Omega I_1^r(q) = -\left(\frac{2}{k^2} - 1 + \Omega^2\right) I_2^r(q),$$
$$I_2^r(q) = \Omega I_c^r \sin\Omega q.$$

Using these identities, we find

$$\mathcal{B}(p,q) := (\mathcal{B}_1(p,q), \mathcal{B}_2(p,q)),$$

where

$$\mathcal{B}_1(p,q) = K_1\pi\lambda\,\mathrm{e}^{-2p} + \frac{2}{3k^2}A\,\mathrm{e}^{-p}\left(2 + k^2(\Omega^2 - 1)\right) I_c^r \sin\Omega q,$$
$$\mathcal{B}_2(p,q) = -\frac{1}{2}I_3^r + \frac{1}{2}B\,\mathrm{e}^{-p}\Omega I_c^r \sin\Omega q.$$

PROPOSITION 4.5. *If K_1 and K_2 are relatively prime positive integers such that $K_1 2\pi/\Omega = K_2 2kK(k)$, then for I_c^r to be nonvanishing it is necessary and sufficient that $K_2 = 1$. In case this condition holds,*

$$I_c^r = 4\left(\frac{\pi^2 K_1}{kK(k)}\right)\frac{\mathbf{q}^{K_1}}{1-\mathbf{q}^{2K_1}} = 4\pi\Omega\frac{\mathbf{q}^{K_1}}{1-\mathbf{q}^{2K_1}},$$

where $\mathbf{q} := \mathrm{e}^{-\pi K'/K}$ is Jacobi's nome.

Proof. The integral I_c^r is computed as in Proposition 4.3 using the Fourier series for $\mathrm{sn}^2(u)$. In fact,

$$I_c^r = -2\int_0^{K_1 2\pi/\Omega} \mathrm{sn}^2(t/k)\cos\Omega t\,dt$$
$$= 4\left(\frac{\pi}{kK}\right)^2 \sum_{n=1}^{\infty}\frac{n\mathbf{q}^n}{1-\mathbf{q}^{2n}}\int_0^{K_1 2\pi/\Omega}\cos\left(\frac{n\pi}{kK}t\right)\cos\Omega t\,dt.$$

After the change of variables $v := \Omega t/K_1$ and substitution from the resonance relation, we obtain

$$I_c^r = 4\left(\frac{\pi}{kK}\right)^2 \frac{kKK_2}{\pi}\sum_{n=1}^{\infty}\frac{n\mathbf{q}^n}{1-\mathbf{q}^{2n}}\int_0^{2\pi}\cos nK_2 v\cos K_1 v\,dv.$$

Thus, I_c^r vanishes unless $nK_2 = K_1$. Since K_1 and K_2 are relatively prime, this means that $I_c^r \neq 0$ exactly when $K_2 = 1$ and K_1 is arbitrary. In this case we obtain

$$I_c^r = 4\left(\frac{\pi^2 K_1}{kK(k)}\right)\frac{\mathbf{q}^{K_1}}{1-\mathbf{q}^{2K_1}}. \qquad \square$$

Finally, we compute I_3^r under the assumption $M(\dot\theta) = m_1 + m_2\dot\theta$. For this we have

$$I_3^r = m_1 \int_0^{K_1 2\pi/\Omega} \dot\theta(t+q)\,dt + m_2 \int_0^{K_1 2\pi/\Omega} \dot\theta^2(t+q)\,dt.$$

In the present case we find, using the resonance relation and the periodicity,

$$\int_0^{K_1 2\pi/\Omega} \dot\theta(t+q)\,dt = -2\pi K_2.$$

Also, as before,

$$\begin{aligned}\int_0^{K_1 2\pi/\Omega} \dot\theta^2(t+q)\,dt &= \int_0^{K_1 2\pi/\Omega} (2\cos(\theta(t+q)) + b^2 - 2)\,dt \\ &= (b^2-2)K_1\frac{2\pi}{\Omega} + 2\int_0^{K_1 2\pi/\Omega} 1 - 2\mathrm{sn}^2[t/k,k]\,dt \\ &= \frac{4}{k}E(\mathrm{am}\,[2K_2K(k),k],k) \\ &= 8K_2\frac{E(k)}{k}.\end{aligned}$$

Thus, we have

$$-\frac{1}{2}I_3^r(q) = m_1\pi K_2 - m_2 4K_2\frac{E(k)}{k}.$$

By the proposition, we see that there are (under appropriate choices of the constant parameters) bifurcating families of periodic solutions for the full model equations at each rotational periodic motion of the gravitationally influenced rotor, whose period is an integer multiple of the natural period of the support oscillator. The fact that the resonances are not restricted to *even* multiples of the period of the support oscillator, as in the case of nonrotational motions, is perhaps expected, since near the separatrix between rotational and nonrotational motions the nonrotational periods are twice as long as the rotational periods. More precisely, if we impose the nondegeneracy condition $K_2 = 1$, then the bifurcation points are the simple solutions of the equations

$$\begin{aligned}\lambda K_1\pi e^{-2p} + \frac{2A}{3k^2}(2 + k^2(\Omega^2-1))4\pi\Omega\frac{\mathbf{q}^{K_1}}{1-\mathbf{q}^{2K_1}}e^{-p}\sin\Omega q = 0,\\ m_1\pi - m_2 4\frac{E(k)}{k} + 2\pi B\Omega^2\frac{\mathbf{q}^{K_1}}{1-\mathbf{q}^{2K_1}}e^{-p}\sin\Omega q = 0.\end{aligned}$$

By eliminating $\sin\Omega q$ from these equations, we find

$$\mathrm{e}^{2p} = \Delta := \frac{3\pi\lambda B\Omega K_1 k^3}{4A(m_1\pi k - m_2 4E(k))(2 + k^2(\Omega^2-1))}.$$

Thus, we can solve for p provided $(m_1\pi k - m_2 4E(k))(2 + k^2(\Omega^2-1)) > 0$. Assuming this condition is satisfied and inserting e^p into the second bifurcation equation, we find $\sin\Omega q := \Lambda$, where

$$\Lambda = -\left(\frac{3\pi\lambda B\Omega K_1 k^3}{4A(m_1\pi k - m_2 4E(k))(2 + k^2(\Omega^2-1))}\right)^{1/2}\left(\frac{m_1\pi k - m_2 4E(k)}{2\pi B\Omega^2 k}\right)\frac{1-\mathbf{q}^{2K_1}}{\mathbf{q}^{K_1}}.$$

Thus, there are two solutions for q provided $-1 < \Lambda < 1$. In addition, it is easy to compute the Jacobian of the two bifurcation equations and deduce that the solutions of the bifurcation equations will both be simple provided $\cos \Omega q \neq 0$. This is as it should be, since the solutions are simple when there are two values of q and not simple at the bifurcation points given by $\sin \Omega q = \pm 1$.

As in the case of the nonrotational motions, if the parameters are fixed, then there are only finitely many resonant motions of the rotor for which the condition $-1 < \Lambda < 1$ is satisfied. This follows as before by showing that Λ is unbounded as $k \to 1$ and $K_1 \to \infty$. Thus, again for rotational motions under a strong gravitational force, infinitely many resonant solutions can lead to nondegenerate first-order bifurcation, but only a finite number of these are excited for a fixed set of parameter values.

We end this section with a useful observation. The divergence of the perturbed vector field, computed in (z, w, x, y)-coordinates, is constant. In fact, the divergence is simply $-\epsilon(\lambda + m_2)$. This is reasonable since λ and m_2 are coefficients of damping in the system. Abel's formula applied to the linear variational equations as in [19, p. 156] implies that the determinant of the linearized Poincaré map is given by

$$\det DP(\xi, \epsilon) = e^{-\epsilon(\lambda + m_2) K_1 2\pi/\Omega}.$$

Thus, the linearized Poincaré map contracts volume and the perturbed periodic solutions found by our bifurcation method are all saddles and sinks. In particular, this shows entrainment (capture) is possible.

4.1. Remarks, experiments, and speculation. We have just shown that there exist choices of the parameters in our model equations such that several periodic solutions, corresponding to rotational motions of the rotor, can coexist. Moreover, these periodic solutions in the four-dimensional phase space are all saddles or sinks. In order to determine the dynamics of the system, we would like further stability information about these periodic solutions. Rigorous stability information may be obtained from a second-order bifurcation analysis. However, we mention that the bifurcating families occur in pairs corresponding to the solutions of the equation $\sin \Omega q = \Lambda$. Generically, one bifurcating family consists of sinks and the other consists of saddles. The basin of attraction of a periodic solution corresponding to a sink is the region in phase space "captured into resonance" or, in other words, it is the entrainment domain. Of course, there is no obvious reason why such a periodic solution will be globally attracting; thus solutions starting outside of the basin of attraction have a different fate. On the other hand, a saddle periodic solution may have a one-, two-, or three-dimensional stable manifold. Solutions starting near the stable manifold may remain near the saddle periodic solution on a very long time scale, appearing to be captured only to leave eventually the vicinity of the saddle periodic solution along its unstable manifold to pass near a second saddle, or perhaps become entrained to a stable periodic solution. If there are several such saddles, this behavior may be very complex.

At the end of the last section we showed that the linearized Poincaré map contracts volume. This fact was used to prove that the perturbed periodic solutions are saddles and sinks. In contrast to the similar analysis of a single forced oscillator, e.g., [19, p. 157] or [9, p. 207], we cannot conclude that there are no invariant closed curves for the *three*-dimensional perturbed Poincaré map. In other words, periodic sinks may coexist with more complicated attractors. Before discussing this possibility more fully, we mention that the analysis completed above only considers the bifurcation of periodic solutions from periodic solutions of the unperturbed oscillators at resonance.

In the example with a strong gravitational force, the mathematical pendulum has, in the phase cylinder, a hyperbolic saddle point corresponding to its unstable equilibrium state, and this rest point has a pair of associated homoclinic trajectories. The dynamics of the perturbed system near the corresponding trajectories in the four-dimensional phase space of the coupled system can perhaps be determined to some extent by analyzing an appropriate "Melnikov" integral. Such an analysis might show the presence of horseshoes. In any case, the existence of complicated attractors remains to be established.

As an excursion in this direction, we have considered a decoupled specialization of our model equations in order to obtain a two-dimensional Poincaré map and the possibility of visual representations of some aspects of the dynamics. To do this, we consider the system

$$\ddot{z} + \Omega^2 z = 0,$$
$$\ddot{\theta} + \sin\theta = -\epsilon(m_1 + m_2\dot{\theta} - Bz\sin\theta).$$

It may be viewed as a single parametrically excited mathematical pendulum.

To study the rotational motions as before, we consider (symplectic) polar coordinates on the phase cylinder to obtain

$$\dot{x} = y(x^2+y^2) + \frac{1}{2}xy(x^2+y^2)^{-3/2} - \frac{\epsilon}{2}\frac{x}{x^2+y^2}h,$$
$$\dot{y} = -x(x^2+y^2) + \frac{1}{2}y^2(x^2+y^2)^{-3/2} - \frac{\epsilon}{2}\frac{y}{x^2+y^2}h,$$

where

$$h = -m_1 + m_2(x^2+y^2) + B\frac{y}{\sqrt{x^2+y^2}}\mathrm{e}^{-p}\cos\Omega t.$$

A comparison of the analysis for the coupled system with the analysis for the single "forced" oscillator as presented in [5], [6] shows that we have already computed the bifurcation function for this system. Here, p is just a parameter, and the scalar bifurcation function is just $\mathcal{B}_2(q)$ as computed above. In fact, for the $(K_1 : K_2)$ resonance, the bifurcation equation is

$$\mathcal{B}_2(q) = \begin{cases} m_1\pi - m_2 4\frac{E(k)}{k} + 2\pi B\Omega^2\frac{\mathbf{q}^{K_1}}{1-\mathbf{q}^{2K_1}}e^{-p}\sin\Omega q = 0 & \text{if } K_2 = 1, \\ K_2(m_1\pi - m_2 4\frac{E(k)}{k}) & \text{if } K_2 \neq 1. \end{cases}$$

In case $K_2 = 1$, the bifurcation equation $\mathcal{B}_2(q) = 0$ is equivalent to $\sin\Omega q := \Lambda$, where

$$\Lambda = -\mathrm{e}^p\left(\frac{m_1\pi k - m_2 4E(k)}{2\pi B\Omega^2 k}\right)\frac{1-\mathbf{q}^{2K_1}}{\mathbf{q}^{K_1}}.$$

Thus, there are two solutions for q provided $-1 < \Lambda < 1$. Here, the linearized Poincaré map is still area contracting (the divergence is $-\epsilon m_2$), so the periodic solutions are again saddles and sinks. However, even in this case we *cannot* conclude that there are no invariant curves in the Poincaré section. This is a result of the fact that, for the rotational motions, the system is defined on an annulus in the phase cylinder whose inner boundary is the separatrix of the unperturbed mathematical pendulum. This fact is reflected in the singularity of the (x, y)-coordinates at $x^2 + y^2 = -\dot{\theta} = 0$. In

FIG. 1. *Schematic representation of basin boundaries for attractors in the perturbed Poincaré map for the rotational motions of the parametrically excited mathematical pendulum. Shaded region is in the basin of attraction of the invariant torus.*

other words, in the (x, y) section, the region corresponding to rotational motion is an annular region surrounding the origin. More precisely, the unperturbed rotational solutions correspond to solutions (outside the separatrices) with energies

$$E = \frac{1}{2}\dot{\theta}^2 + 1 - \cos\theta$$
$$= \frac{1}{2}(x^2 + y^2)^2 + 1 - \frac{x}{\sqrt{x^2 + y^2}} > 1.$$

When the system is perturbed, solutions can cross into the region with $E < 1$ and then eventually cross the curve $\dot{\theta} = 0$, where the vector field is singular. Thus, the area of the region corresponding to rotational motions is not preserved. Of course, the fact that the linearized Poincaré map is area contracting does imply that there is at most one invariant curve. If there were two invariant curves, the annular region bounded by these curves would be invariant. Numerical experiments suggest, in fact, that invariant curves exit. This suggests that a similar phenomenon is possible for the coupled system, but at present we do not know how to examine this possibility rigorously.

We have investigated the dynamics of the uncoupled system in the region of parameter space corresponding to parameter values where the *coupled* system has periodic solutions arising from the bifurcation theory given previously. A useful example is provided by the following choice of parameters:

$$K_1 = 1, \quad \Omega = 4, \quad A = 4, \ \lambda = 0, \quad B = 4, \quad m_1 = 10$$

with m_2 and p variable.

Let Γ denote the $(K_1 : K_2) = (1 : 1)$ resonant periodic solution of the mathematical pendulum. This solution is given by the elliptic modulus $k_* \approx 0.47$. In the original coordinates, it is the solution starting at $(\theta, \dot{\theta}) = (0, -2/k_*)$. Recall that a

necessary condition for the bifurcation equations obtained for the *coupled* system to have solutions in this case is $m_1\pi k_* - m_2 4E(k_*) > 0$. Thus, for the given parameters, we must have $0 < m_2 < 5\pi k_*/(2E(k_*)) \approx 2.495$. If this condition is satisfied, p is determined by the previously given formula $e^{2p} = \Delta$.

The unperturbed Poincaré map for the uncoupled system giving the return to the (x, y) plane after time $2\pi/\Omega$ has Γ as an invariant curve. In fact, the unperturbed Poincaré map is the identity on Γ. In addition, for small positive ϵ, we have proved that there are two fixed points for the perturbed Poincaré map near the zeros of $\mathcal{B}_2(q)$, and these fixed points correspond to the persistent periodic solutions. For the resonances with $K_2 \neq 1$, the bifurcation function reduces to $b(k) := K_2(m_1\pi - m_2 4E(k)/k)$ and is independent of q. It is easy to see that the dense set of resonant orbits such that $K_1/K_2 < 1$ lies "outside" Γ in the Poincaré section, while the dense set of resonant orbits such that $K_2 \neq 1$ and $K_1/K_2 > 1$ lies "inside" Γ. Moreover, the reduced bifurcation function b is positive inside and on Γ. That is, there is an unperturbed periodic solution Γ_0 of the mathematical pendulum corresponding to some $k_0(m_2) < k_*$ such that $b(k_0) = 0$. In particular, Γ_0 surrounds Γ, $b(k) < 0$ for all $k < k_0$, and the resonant orbits outside Γ_0 correspond to resonances such that $K_1/K_2 < 1$. Thus, for these resonant orbits $K_2 \neq 1$ and $\mathcal{B}_2 = b$.

In general, some of the resonant orbits corresponding to $K_2 = 1$ and $K_1 > 1$ can be excited and additional periodic solutions can occur. But for our choice of parameters this does not happen. Thus, in a manner similar to the discussion in [19, pp. 161–175], we observe that perturbed trajectories of the Poincaré map tend to drift outward toward Γ_0 from the region inside Γ_0, except for the resonance layer near Γ, and they tend to drift inward toward Γ_0 from the outside. In particular, there are no periodic solutions excited by the perturbation except for those on Γ.

The existence of a periodic sink and a periodic saddle for the perturbed Poincaré map corresponding to the perturbed periodic solutions is consistent with these facts. This is exactly the situation observed in numerical simulation. In addition, the positions of the bifurcation points as predicted independently by solving the bifurcation equations $e^{2p} = \Delta$ and $\sin \Omega q = \Lambda$ are also confirmed. However, the facts about the sign of b and the implied drift directions for the perturbed solutions indicate that there is also the possibility of a nonperiodic attractor Γ_ϵ near Γ_0 coexistent with the periodic sink. Our numerical experiments confirm the existence of such an invariant attracting set. It appears to be a smooth curve for $0 < m_2 < 2.459$ and all sufficiently small $\epsilon > 0$.

From the discussion above, the periodic sink lies inside the region bounded by this invariant curve. Because there are two attractors, the entrainment domain (the basin of attraction of the periodic sink) shares a common boundary with the basin of attraction of the invariant curve. Figure 1 schematically shows the basins of attraction for the two attractors. Aside from the fact that the spiral basin of attraction of the periodic sink is very thin for small ϵ, we also see that solutions with initial values "outside" Γ_ϵ are never entrained to the periodic sink. We expect the entrainment domain for the coupled system to be at least as complex.

REFERENCES

[1] A. A. ANDRONOV, E. A. VITT, AND S. E. KHAIKEN, *Theory of Oscillators,* Pergamon Press, Oxford, 1966.

[2] V. I. ARNOLD, *Geometric Methods in the Theory of Ordinary Differential Equations,* Springer-Verlag, New York, 1982.

[3] V. I. ARNOLD, *Loss of stability of self-oscillations close to resonance and versal deformations of equivariant vector fields,* Functional Anal. Appl., 11 (1977), pp. 1–10.
[4] P. F. BYRD AND M. D. FRIEDMAN, *Handbook of Elliptic Integrals for Engineers and Scientists,* 2nd ed., Springer-Verlag, Berlin, 1971.
[5] C. CHICONE, *Bifurcation of nonlinear oscillations and frequency entrainment near resonance,* SIAM J. Math. Anal., 23 (1992), pp. 1577–1608.
[6] ———, *Lyapunov-Schmidt reduction and Melnikov integrals for bifurcation of periodic solutions in coupled oscillators,* J. Differential Equations, 112 (1994), pp. 407–447.
[7] S. N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory,* Springer-Verlag, New York, 1982.
[8] S. P. DILIBERTO, *On systems of ordinary differential equations,* in Contributions to the Theory of Nonlinear Oscillations, Ann. of Math. Stud., Vol. 20, Princeton University Press, Princeton, 1950, pp. 1–38.
[9] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields,* 2nd ed., Springer-Verlag, New York, 1986.
[10] C. HAYASHI, *Nonlinear Oscillations in Physical Systems,* McGraw–Hill, New York, 1964.
[11] V. K. MELNIKOV, *On the stability of the center for time periodic perturbations,* Trans. Moscow Math. Soc., 12 (1963), pp. 1–57.
[12] N. MINORSKY, *Nonlinear Oscillations,* Van Nostrand, Princeton, 1962.
[13] A. NEISHTADT, *Averaging, capture into resonance, and chaos in nonlinear systems,* in Chaos/XAOC, Soviet–American Perspectives on Nonlinear Science, D. K. Campbell, ed., American Institute of Physics, New York, 1990.
[14] R. H. RAND AND P. J. HOLMES, *Bifurcation of periodic motions in two weakly coupled van der Pol oscillators,* Internat. J. Non-Linear Mech., 15 (1980), pp. 387–399.
[15] R. H. RAND, R. J. KINSEY, AND D. L. MINGORI, *Dynamics of spinup through resonance,* Internat. J. Non-Linear Mech., 27 (1992), pp. 489–502.
[16] J. A. SANDERS AND F. VERHULST, *Averaging Methods in Nonlinear Dynamical Systems,* Springer-Verlag, New York, 1985.
[17] J. J. STOKER, *Nonlinear Vibrations,* John Wiley, New York, 1950.
[18] S. W. WIGGINS, *Global Bifurcations and Chaos,* Springer-Verlag, New York, 1988.
[19] ———, *Introduction to Applied Nonlinear Dynamical Systems and Chaos,* Springer-Verlag, New York, 1990.
[20] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis,* American ed., Macmillan, New York, 1946.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1284–1305, September 1995

011

ON THE BIFURCATION STRUCTURE OF NONLINEAR PERTURBATIONS OF HILL'S EQUATIONS AT BOUNDARY POINTS OF THE CONTINUOUS SPECTRUM*

TASSILO KÜPPER† AND THOMAS MRZIGLOD†

Abstract. Nonlinear perturbations of Hill's equations have been studied as a first application of a general operator-theoretic approach for treating bifurcation at boundary points of the continuous spectrum. It has been established that there is bifurcation into the gap at distinguished boundary points of the spectrum; moreover, for fixed parameters in the gap there are m distinct solutions where m can be characterized by the number of negative eigenvalues of an associated linear eigenvalue problem. For a class of nonlinear Hill equations with a nonlinearity concentrated on a finite inverval $[-N, N]$, we are able to reduce the problem to an auxiliary nonlinear Sturm–Liouville problem with parameter dependent boundary conditions. The reduction is based on the knowledge of the stable/unstable spaces of the linearized problem. Although the reduced problem is of a complicated nature, we can analyze its bifurcation structure by a modified Lyapunov–Schmidt procedure. In that way we provide a detailed analysis of both the reduced and the original problem and we can explain various phenomena which occur in connection with bifurcation from the continuous spectrum. In particular, we detect global effects of the presence of continuous spectrum and we provide a mechanism to understand results on the various numbers of solutions.

Key words. bifurcation from continuous spectrum, gap bifurcation, nonlinear Hill equation

AMS subject classifications. 34A47, 34B15

1. Introduction.

In this paper we study in detail the bifurcation behavior of a special class of nonlinear Hill equations

$$-\left(p(x)u'(x)\right)' + q(x)u(x) + f\left(x, u(x)\right)u(x) = \lambda s(x)u(x) \quad (x \in \mathbb{R}), \tag{1.1}$$

$$u \in L^p(\mathbb{R}) \quad (1 \le p < \infty), \tag{1.2}$$

where p, q, s, and f statisfy the following conditions:

(H0) $p, p', q, s : \mathbb{R} \to \mathbb{R}$ are bounded, piecewise continuous, and two-periodic on $\mathbb{R}$; in addition,

$$0 < p_0 \le p(x), \quad 0 < s_0 \le s(x) \quad (x \in \mathbb{R}).$$

(H1) $f : \mathbb{R} \times [-y_0, y_0] \to \mathbb{R}$ $(y_0 > 0)$ is such that
- (i) $f(x, y) = 0$ if $x \notin [-N, N]$ for some $N \in \mathbb{N}$,
- (ii) $f(\cdot, y) : \mathbb{R} \to \mathbb{R}$ is piecewise continuous and bounded,
- (iii) $f(x, \cdot) : [-y_0, y_0] \to \mathbb{R}$ is continuous, $f(x, 0) = 0$.

(H2) $f(x, \cdot) \in C^1[-y_0, y_0]$, $\frac{\partial f}{\partial y} : [-N, N] \times [-y_0, y_0] \to \mathbb{R}$ is bounded.

According to the hypotheses on the coefficients, we are looking for solutions $u \in C^1(\mathbb{R}) \cap L^p(\mathbb{R})$ with a piecewise continuous second derivative.

* Received by the editors June 23, 1993; accepted for publication February 8, 1994.
† Mathematisches Institut der Universität zu Köln, D-50923 Köln, Germany.

Nonlinear Hill equations have been studied as a first application [13] of a general functional analytical approach [12] for treating bifurcation at boundary points of the continuous spectrum for operator equations of the form

$$Lu + \mathcal{N}(u) = \lambda u;$$

here L denotes a linear self-adjoint operator with spectrum $\sigma(L)$, and the nonlinearity $\mathcal{N}$ is taken as a gradient operator $\mathcal{N} = \varphi'$ such that $\mathcal{N}(0) = 0$, $\mathcal{N}'(0) = 0$. Using variational methods [4]–[15], [24] it has been shown that there is indeed bifurcation at distinguished boundary points of $\sigma(L)$ into the gaps under additional monotonicity hypotheses on $\mathcal{N}$. On the other hand, the existence of a specific number of different solutions for each $\lambda \notin \sigma(L)$ has been established as well [1], [2], [8]; here the number of solutions depends on the number of negative eigenvalues of an auxiliary linear eigenvalue problem. In addition, both bifurcation from 0 and asymptotic bifurcation from ∞ have been found.

For the special class of differential operators with a nonlinearity concentrated on a finite interval $[-N, N]$, we are able to reduce the problem to an auxiliary Sturm–Liouville problem on the interval $[-N, N]$, where the boundary conditions depend in a complicated way on the parameter λ. The reduction is based on the knowledge of the unstable (resp., stable) spaces of the linearized problem; since the problem is linear outside $[-N, N]$ these "manifolds" are explicitly available in terms of a fundamental system. Although the reduced problem is of a complicated nature, it can nevertheless be analysed by classical methods. In this way we are able to provide a detailed analysis of the bifurcation behavior of both the reduced and the original problem, and we can explain various phenomena which occur in connection with bifurcation from the continuous spectrum by properties of the underlying reduced problem. In particular, we understand the

(i) background of bifurcation at boundary points of $\sigma(L)$;

(ii) coexistence of bifurcation from 0 and asymptotic bifurcation from ∞;

(iii) results by Alama and Li [2] and Heinz [8] on the number of "branches" over the gaps, in the context of bifurcation from possibly different endpoints of the spectrum which disappear over the spectrum and return over the gaps;

(iv) global effects of the presence of a continuous spectrum on the existence of nontrivial solutions. We recall that global effects of the continuous spectrum have already been addressed by Stuart [19], [20] in an extension of the classical bifurcation result by Rabinowitz [17]. A detailed study of the global bifurcation behavior for this kind of problem is in preparation (Mrziglod [16]).

Since our approach is not based on variational methods, we are also able to drop the usual monotonicity requirements on $\mathcal{N}$; instead we can work out explicit criteria for bifurcation, even for nonlinearities which are not covered by the results obtained so far. For example, we can treat nonlinearities which are even or change sign. Through a thorough analysis of these problems, a series of so far hidden relations is uncovered which play an important role in the understanding of bifurcation from the continuous spectrum. Although the concentration of the nonlinearity on a finite interval appears as a severe restriction, we guess that our results are of a general nature and are able to shed a new light onto the phenomenon of bifurcation from the continuous spectrum. We expect that this approach can be extended to general nonlinearities by replacing the stable/unstable spaces of the linear problem outside $[-N, N]$ by the corresponding manifolds. Since its performance will involve a lot of technical details, we consider it as an interesting project for future research.

The paper is organized in the following way: In §2 basic facts for the linear Hill equation are collected; §3 contains the derivation of the reduced problem, which is analyzed in §4. In §5 we return to the original problem and discuss the results in terms of the original problem.

2. Properties of the linearized equation. To understand the local bifurcations from the trivial solution we recollect some properties of the linear equation

$$-(ph')' + (q - \lambda s)h = 0. \tag{2.1}$$

According to Floquet's theory (Eastham [3]) we choose a fundamental system of solutions $\varphi_1(x,\lambda)$ and $\varphi_2(x,\lambda)$ determined by the initial conditions

$$\begin{aligned} \varphi_1(-N,\lambda) &= 1, \qquad \varphi_1'(-N,\lambda) = 0, \\ \varphi_2(-N,\lambda) &= 0, \qquad \varphi_2'(-N,\lambda) = 1. \end{aligned} \tag{2.2}$$

For $D(\lambda) := \varphi_1(N,\lambda) + \varphi_2'(N,\lambda)$ there are two roots ϱ_1, ϱ_2 of the characteristic equation

$$\varrho^2 - D(\lambda)\varrho + 1 = 0$$

satisfying $\varrho_1 \cdot \varrho_2 = 1$. Furthermore, there are corresponding solutions w_1, w_2 of (2.1) satisfying $w_i(x+2N) = \varrho_i w_i(x)$. Recall that $\varphi_1(N,\lambda), \varphi_2(N,\lambda), D(\lambda)$ are analytic in λ. To simplify the notation, throughout the rest of the paper we set

$$\varphi_i(\lambda) := \varphi_i(N,\lambda), \ \varphi_i'(\lambda) := \varphi_i'(N,\lambda) \quad (i = 1,2).$$

If the differential equation (2.1) is associated with

1. periodic boundary conditions

$$h(N) = h(-N), \qquad h'(N) = h'(-N), \tag{2.3}$$

there are infinitely many eigenvalues $\lambda_0 < \lambda_1 \le \lambda_2 < \lambda_3 \le \cdots$;

2. semiperiodic boundary conditions

$$h(N) = -h(-N), \qquad h'(N) = -h'(-N), \tag{2.4}$$

there are infinitely many eigenvalues $\mu_0 \le \mu_1 \le \cdots$.

They satisfy the relation

$$\lambda_0 < \mu_0 \le \mu_1 < \lambda_1 \le \lambda_2 < \mu_2 \le \mu_3 < \lambda_3 \le \cdots.$$

These eigenvalues, the range of $D(\lambda)$, and the spectrum $\sigma(L)$ of the differential equation (2.1) on $L^2(\mathbb{R})$ are related in the following way:

$$\begin{aligned} \sigma = \sigma(L) &= \bigcup_{j=0}^{\infty} [\lambda_{2j}, \mu_{2j}] \ \cup \ \bigcup_{j=0}^{\infty} [\mu_{2j+1}, \lambda_{2j+1}] \\ &= \{\lambda \in \mathbb{R} / \ |D(\lambda)| \le 2\}. \end{aligned}$$

The boundary points of σ are precisely the simple eigenvalues of (2.1) with (2.3) or (2.4); in each boundary point λ_0 of σ the condition either $\varphi_2(\lambda_0) \neq 0$ or $\varphi_1'(\lambda_0) \neq 0$, and $D'(\lambda_0) \neq 0$ holds.

For fixed λ in each of the intervals

$$(-\infty, \lambda_0), [\lambda_0, \mu_0], (\mu_0, \mu_1), \ldots,$$

there is a specific fundamental system characterized by its behavior for $x \to \pm\infty$. For λ in a gap of the spectrum (i.e., $|D(\lambda)| > 2$) there is a special fundamental system u_{-1}, u_1 such that $u_i(x) \to 0 (x \to i\infty)$ and $u_i(x)$ is unbounded for $x \to -i\infty$ $(i = -1, 1)$.

For $|D(\lambda)| \geq 2$ set $B(\lambda) := \operatorname{sgn}(D(\lambda))\sqrt{D^2(\lambda) - 4}$. Then the functions u_{-1}, u_1 are explicitly given in the following cases:

(i) $\varphi_2(\lambda) \neq 0$,

$$u_i(x,\lambda) := 2\varphi_2(\lambda)\varphi_1(x,\lambda) - [(\varphi_1(\lambda) - \varphi_2'(\lambda)) + iB(\lambda)]\,\varphi_2(x,\lambda); \tag{2.5}$$

(ii) $\varphi_2(\lambda) = 0$ and $\varphi_1(\lambda) \neq 0$,

$$u_i(x,\lambda) := [-i\,(\varphi_1(\lambda) - \varphi_2'(\lambda)) + B(\lambda)]\,\varphi_1(x,\lambda) - 2i\varphi_1'(\lambda)\varphi_2(x,\lambda); \tag{2.6}$$

(iii) $\varphi_2(\lambda) = \varphi_1'(\lambda) = 0$ and $\operatorname{sgn}D(\lambda) = \operatorname{sgn}(\varphi_1(\lambda) - \varphi_2'(\lambda))$,

$$u_{-1}(x,\lambda) := \varphi_1(x,\lambda),\ u_1(x,\lambda) := \varphi_2(x,\lambda); \tag{2.7}$$

(iv) $\varphi_2(\lambda) = \varphi_1'(\lambda) = 0$ and $\operatorname{sgn}D(\lambda) \neq \operatorname{sgn}(\varphi_1(\lambda) - \varphi_2'(\lambda))$,

$$u_{-1}(x,\lambda) := \varphi_2(x,\lambda),\ u_1(x,\lambda) := \varphi_1(x,\lambda). \tag{2.8}$$

For all $x \in \mathbb{R}$ the functions u_{-1}, u_1 satisfy

$$\begin{aligned} u_{-1}(x - 2N, \lambda) &= \varrho(\lambda)\ u_{-1}(x,\lambda), \\ u_1(x + 2N, \lambda) &= \varrho(\lambda)\ u_1(x,\lambda), \end{aligned} \tag{2.9}$$

where

$$\varrho(\lambda) := \frac{1}{2}\left[D(\lambda) - \operatorname{sgn}(D(\lambda))\sqrt{D^2(\lambda) - 4}\right] \tag{2.10}$$

is the characteristic root determined by $|\varrho(\lambda)| \leq 1$.

LEMMA 2.1. *Suppose* (H0) *holds and assume that* $w_{-1} \in L^p(-\infty, -N)$, $w_1 \in L^p(N, \infty)$ *are solutions of the differential equation* (2.1).

(i) *If* $\lambda \in \sigma$ *then* $w_{-1} \equiv 0$, $w_1 \equiv 0$.

(ii) *If* $\lambda \notin \sigma$ *then, for some constants* c_{-1}, $c_1 \in \mathbb{R}$,

$$w_{-1} = c_{-1}u_{-1|(-\infty,-N)}, \quad w_1 = c_1 u_{1|(N,\infty)}. \tag{2.11}$$

Proof. Part (ii) follows immediately from property (2.9). Part (i) can be derived from the existence of two linearly independent (complex-valued) solutions u_{-1}, u_1 of (2.1) satisfying $|u_i(x + 2N)| = |u_i(x)|$ for all $x \in \mathbb{R}$. □

3. The reduced problem. If $u(x,\lambda)$ is a solution of the differential equation (1.1) we can split it as follows:

$$u(x,\lambda) = \begin{cases} v_{-1}(x,\lambda) & (x \leq -N), \\ v(x,\lambda) & (-N \leq x \leq N), \\ v_1(x,\lambda) & (x \geq N). \end{cases} \tag{3.1}$$

Because of the restrictions on the support of f, the functions v_{-1} and v_1 solve the linear differential equation (2.1); hence, if $u \in L^p(\mathbb{R})$ and $|D(\lambda)| > 2$, there are constants c_{-1}, c_1 such that

$$v_{-1}(x,\lambda) = c_{-1}u_{-1}(x,\lambda), \qquad v_1(x,\lambda) = c_1 u_1(x,\lambda).$$

The function $v(x,\lambda)$ is a solution of the nonlinear differential equation on $[-N,N]$; for smoothness it satisfies the boundary conditions

$$\begin{aligned} v(-N,\lambda) &= v_{-1}(-N,\lambda) = c_{-1}u_{-1}(-N,\lambda), \\ v'(-N,\lambda) &= v'_{-1}(-N,\lambda) = c_{-1}u'_{-1}(-N,\lambda), \end{aligned} \tag{3.2}$$

$$\begin{aligned} v(N,\lambda) &= v_1(N,\lambda) = c_1 u_1(N,\lambda), \\ v'(N,\lambda) &= v'_1(N,\lambda) = c_1 u'_1(N,\lambda). \end{aligned} \tag{3.3}$$

If $u \not\equiv 0$ then $|u_1(-N,\lambda)| + |u'_{-1}(-N,\lambda)| > 0$, $|u_1(N,\lambda)| + |u'_1(N,\lambda)| > 0$, and the unknown parameters c_{-1} and c_1 can be eliminated, leading to the following Sturm–Liouville boundary conditions for v:

$$\begin{aligned} u'_{-1}(-N,\lambda)v(-N,\lambda) - u_{-1}(-N,\lambda)v'(-N,\lambda) &= 0, \\ u'_1(N,\lambda)v(N,\lambda) - u_1(N,\lambda)v'(N,\lambda) &= 0. \end{aligned}$$

Set

$$\alpha_i(\lambda) := -iu'_i(-N,\lambda),\ \beta_i(\lambda) := u_i(-N,\lambda) \quad (i = -1, 1). \tag{3.4}$$

Using (2.9) in the coefficients of the second boundary condition leads to the Sturm–Liouville problem on $[-N,N]$:

$$-(pv')' + qv + f(x, v(x))\, v = \lambda s v, \tag{3.5}$$

$$\begin{aligned} \alpha_{-1}(\lambda)v(-N) - \beta_{-1}(\lambda)v'(-N) &= 0, \\ \alpha_1(\lambda)v(N) + \beta_1(\lambda)v'(N) &= 0. \end{aligned} \tag{3.6}$$

The relation between solutions of the problem (1.1), (1.2) and the reduced system (3.5), (3.6) is precisely described in the following theorem.

THEOREM 3.1. *Let* (H0) *and* (H1) *hold.*

(i) *Nontrivial L^p solutions of equation* (1.1) *exist only if $\lambda \notin \sigma$, i.e., $|D(\lambda)| > 2$.*

(ii) *Suppose $\lambda \notin \sigma$.*

(a) *If $u(x,\lambda) \in L^p(\mathbb{R})$ is a nontrivial solution of* (1.1) *then $v_\lambda := u(\cdot,\lambda)_{|[-N,N]}$ is a nontrivial solution of the Sturm–Liouville problem* (3.5), (3.6).

(b) *If v_λ is a nontrivial solution of the Sturm–Liouville problem* (3.5), (3.6) *then the function*

$$u(x,\lambda) := \begin{cases} c_{-1}(\lambda)u_{-1}(x,\lambda) & (x < -N), \\ v_\lambda(x) & (-N \le x \le N), \\ c_1(\lambda)u_1(x-2N,\lambda) & (x > N) \end{cases} \tag{3.7}$$

is a nontrivial solution of (1.1) *in $L^p(\mathbb{R})$; here the coefficients $c_i(\lambda)$ are given by*

$$c_i(\lambda) := \begin{cases} v_\lambda(iN)/u_i(-N,\lambda) & \text{if } u_i(-N,\lambda) \neq 0, \\ v'_\lambda(iN)/u'_i(-N,\lambda) & \text{otherwise.} \end{cases} \tag{3.8}$$

Proof. By the Gronwall inequality and $f(x,0) = 0$, every solution v of (3.5) which has a double zero is $v \equiv 0$. Hence part (i) follows by Lemma 2.1. To prove (ii(b)) one has to show the continuity of u and u' defined by (3.7) at $x = \pm N$. □

Using the multiplier property (2.9) for solutions of the linear problem, we can set up a relation between the norms of u_λ and v_λ which will be useful for the interpretation of our results in later sections.

LEMMA 3.2. *Suppose that* (H0) *and* (H1) *hold and let* $(v_\lambda, \lambda)(\lambda \notin \sigma)$ *be a solution of* (3.5), (3.6) *with* $\alpha_i(\lambda)$, $\beta_i(\lambda)$ *as defined in* (3.4). *Then the norm of the corresponding solution* (u_λ, λ) *given by* (3.7) *can be expressed in the following way:*

$$\|u_\lambda\|_p^p = \int_{-N}^{N} |v_\lambda(x)|^p dx + \frac{1}{1-|\varrho(\lambda)|^p}\left[|c_{-1}(\lambda)|^p\ |\varrho(\lambda)|^p \int_{-N}^{N} |u_{-1}(x,\lambda)|^p dx + |c_1(\lambda)|^p \int_{-N}^{N} |u_1(x,\lambda)|^p dx\right],$$

where $\varrho(\lambda)$ *is given by* (2.10) *and* $|\varrho(\lambda)| < 1$.

Proof. By Theorem 3.1 we have

$$\begin{aligned}\|u_\lambda\|_p^p &= \int_{-N}^{N}|v_\lambda(x)|^p dx + |c_{-1}(\lambda)|^p \int_{-\infty}^{-N}|u_{-1}(x,\lambda)|^p dx + |c_1(\lambda)|^p \int_{N}^{\infty}|u_1(x-2N,\lambda)|^p dx \\ &= \int_{-N}^{N}|v_\lambda(x)|^p dx + |c_{-1}(\lambda)|^p|\varrho(\lambda)|^p \int_{-\infty}^{N}|u_{-1}(x,\lambda)|^p dx + |c_1(\lambda)|^p \int_{-N}^{\infty}|u_1(x,\lambda)|^p dx.\end{aligned}$$

Using (2.9) we obtain

$$\begin{aligned}\int_{-N}^{\infty}|u_1(x,\lambda)|^p dx &= \sum_{k=0}^{\infty}\int_{-N}^{N}|u_1(x+2kN,\lambda)|^p dx \\ &= \sum_{k=0}^{\infty}|\varrho(\lambda)|^{kp}\int_{-N}^{N}|u_1(x,\lambda)|^p dx \\ &= \frac{1}{1-|\varrho(\lambda)|^p}\int_{-N}^{N}|u_1(x,\lambda)|^p dx\end{aligned}$$

and, similarly,

$$\int_{-\infty}^{N}|u_{-1}(x,\lambda)|^p dx = \frac{1}{1-|\varrho(\lambda)|^p}\int_{-N}^{N}|u_{-1}(x,\lambda)|^p dx. \qquad \square$$

4. Bifurcation analysis of the reduced problem. Here we will study the reduced problem as a bifurcation problem. We assume that λ_0 is a boundary point of σ, i.e., $\lambda_0 \in \partial\sigma$ or $D^2(\lambda_0) = 4$ and $D'(\lambda_0) \neq 0$. We will continue the coefficients $\alpha_i(\lambda)$, $\beta_i(\lambda)$ $(i = -1, 1)$ up to λ_0 in such a way that λ_0 becomes a bifurcation point for a suitably modified problem, and we will also see that no value of λ with $|D(\lambda)| > 2$ can be a bifurcation point.

Firstly we consider the linear boundary value problem

$$-(ph')' + qh = \lambda sh, \tag{4.1}$$

$$\begin{aligned} \alpha_{-1}(\lambda)h(-N) - \beta_{-1}(\lambda)h'(-N) &= 0, \\ \alpha_1(\lambda)h(N) + \beta_1(\lambda)h'(N) &= 0. \end{aligned} \tag{4.2}$$

LEMMA 4.1. *Suppose that* (H0) *holds and the coefficients* α_i, β_i *are defined as in* (3.4).

(i) *If* $\lambda \notin \sigma$ (*i.e.,* $D^2(\lambda) > 4$) *then* $h = 0$ *is the only solution of* (4.1), (4.2).

(ii) *Assume that* $\lambda = \lambda_0 \in \partial\sigma$.

(a) *If* $\varphi_2(\lambda_0) \neq 0$, *there exists* $\delta > 0$ *such that* $\varphi_2(\lambda) \neq 0$ *for* $\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta)\backslash\sigma$; *the coefficients* $\alpha_i(\lambda), \beta_i(\lambda)$ $(i = -1, 1)$ *can be continued up to* λ_0 *such that*

$$\begin{aligned} \alpha_i(\lambda_0) &:= \lim_{\substack{\lambda \to \lambda_0 \\ \lambda \notin \sigma}} \alpha_i(\lambda) = i\left(\varphi_1(\lambda_0) - \varphi_2'(\lambda_0)\right), \\ \beta_i(\lambda_0) &:= \lim_{\substack{\lambda \to \lambda_0 \\ \lambda \notin \sigma}} \beta_i(\lambda) = 2\varphi_2(\lambda_0). \end{aligned}$$

(b) *If* $\varphi_2(\lambda_0) = 0$, *there exists* $\delta > 0$ *such that* $\varphi_1'(\lambda) \neq 0$ *for* $\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta)\backslash\sigma$; *the coefficients* $\alpha_i(\lambda), \beta_i(\lambda)(i = -1, 1)$ *can be continued up to* λ_0 *such that*

$$\begin{aligned} \alpha_i(\lambda_0) &:= \lim_{\substack{\lambda \to \lambda_0 \\ \lambda \notin \sigma}} \alpha_i(\lambda) = 2\varphi_1'(\lambda_0), \\ \beta_i(\lambda_0) &:= \lim_{\substack{\lambda \to \lambda_0 \\ \lambda \notin \lambda_0}} \beta_i(\lambda) = -i\left(\varphi_1(\lambda_0) - \varphi_2'(\lambda_0)\right). \end{aligned}$$

The function

$$\psi(x) := \beta_{-1}(\lambda_0)\varphi_1(x, \lambda_0) + \alpha_{-1}(\lambda_0)\varphi_2(x, \lambda_0) \tag{4.3}$$

is a nontrivial solution of (4.1), (4.2) *for* $\lambda = \lambda_0$. *If* $D(\lambda_0) = 2$, ψ *has period* $2N$; *if* $D(\lambda_0) = -2$, ψ *has period* $4N$. *Furthermore, either* $(\lambda_0 - \delta, \lambda_0) \cap \sigma = \emptyset$ *or* $(\lambda_0, \lambda_0 + \delta) \cap \sigma = \emptyset$.

Proof. (i) Suppose h is a solution of (4.1), (4.2) for $\lambda \notin \sigma$. By Theorem 3.1 there exists a solution $u_\lambda \in L^p(\mathbb{R})$ of (1.1) for $f \equiv 0$ with $u_{\lambda|[-N,N]} = h$, hence u_λ is a solution of (2.1) too. Since the spectrum of (2.1) is purely continous, $u_\lambda \equiv 0$ and $h \equiv 0$.

(ii) Since $\varphi_2(\lambda)$ is a continuous function of λ, there exists a $\delta > 0$ such that $\varphi_2(\lambda) \neq 0$ for $\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta)\backslash\sigma$ in case (a). If $\varphi_2(\lambda_0) = 0$, then $\varphi_1'(\lambda_0) \neq 0$ since $\lambda_0 \in \partial\sigma$, and, similarly, one obtains $\varphi_1'(\lambda) \neq 0$ for $\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta)\backslash\sigma$ for some $\delta > 0$.

Now consider the function ψ as defined in (4.3). In both cases we have $\beta_{-1}^2(\lambda_0) + \alpha_{-1}^2(\lambda_0) \neq 0$, so ψ is a nontrivial solution of (4.1) and

$$\begin{aligned} \psi(-N) &= \beta_{-1}(\lambda_0)\varphi_1(-N, \lambda_0) + \alpha_{-1}(\lambda_0)\varphi_2(-N, \lambda_0) = \beta_{-1}(\lambda_0), \\ \psi'(-N) &= \beta_{-1}(\lambda_0)\varphi_1'(-N, \lambda_0) + \alpha_{-1}(\lambda_0)\varphi_2'(-N, \lambda_0) = \alpha_{-1}(\lambda_0), \end{aligned}$$

hence ψ satisfies the boundary condition (4.2) at $-N$.

At λ_0 we have $D(\lambda_0) = 2j$ with $j = 1$ or $j = -1$. In case (a) it follows that $\psi(-N) = \beta_{-1}(\lambda_0) = 2\varphi_2(\lambda_0)$ and

$$\begin{aligned} \psi(N) &= \beta_{-1}(\lambda_0)\varphi_1(\lambda_0) + \alpha_{-1}(\lambda_0)\varphi_2'(\lambda_0) \\ &= 2\varphi_2(\lambda_0)\varphi_1(\lambda_0) - (\varphi_1(\lambda_0) - \varphi_2'(\lambda_0))\varphi_2(\lambda_0) \\ &= \varphi_2(\lambda_0)\varphi_1(\lambda_0) + \varphi_2'(\lambda_0)\varphi_2(\lambda_0) \\ &= D(\lambda_0)\varphi_2(\lambda_0) \\ &= 2j\varphi_2(\lambda_0) = j\psi(-N), \end{aligned}$$

and using $\varphi_1(\lambda_0)\varphi_2'(\lambda_0) - \varphi_1'(\lambda_0)\varphi_2(\lambda_0) = 1$ we get

$$\begin{aligned}\psi'(N) &= \beta_{-1}(\lambda_0)\varphi_1'(\lambda_0) + \alpha_{-1}(\lambda_0)\varphi_2'(\lambda_0)\\ &= 2\varphi_2(\lambda_0)\varphi_1'(\lambda_0) - (\varphi_1(\lambda_0) - \varphi_2'(\lambda_0))\varphi_2'(\lambda_0)\\ &= -2 + 2\varphi_2'(\lambda_0)\varphi_1(\lambda_0) - (\varphi_1(\lambda_0) - \varphi_2'(\lambda_0))\varphi_2'(\lambda_0)\\ &= -\tfrac{1}{2}D^2(\lambda_0) + (\varphi_1(\lambda_0) + \varphi_2'(\lambda_0))\varphi_2'(\lambda_0)\\ &= -\tfrac{1}{2}(\varphi_1(\lambda_0) + \varphi_2'(\lambda_0))^2 + (\varphi_1(\lambda_0) + \varphi_2'(\lambda_0))\varphi_2'(\lambda_0)\\ &= -\tfrac{1}{2}(\varphi_1(\lambda_0) + \varphi_2'(\lambda_0))(\varphi_1(\lambda_0) - \varphi_2'(\lambda_0))\\ &= -j(\varphi_1(\lambda_0) - \varphi_2'(\lambda_0)) = j\psi'(-N).\end{aligned}$$

Hence, ψ is $2N$-periodic ($j = 1$) (resp., $2N$-antiperiodic ($j = -1$)) and, by $\beta_{-1}(\lambda_0) = \beta_1(\lambda_0)$ and $\alpha_{-1}(\lambda_0) = -\alpha_1(\lambda_0)$, a solution of (4.1), (4.2).

Case (b) follows similarly. □

For a standard investigation of the bifurcation problem using the implicit function theorem, the functions $\alpha_i(\lambda), \beta_i(\lambda)$ need to be differentiated with respect to λ. Since $\lim_{\lambda\to\lambda_0} |\alpha_i'(\lambda)| = \infty$ in case (a) and $\lim_{\lambda\to\lambda_0} |\beta_i'(\lambda)| = \infty$ in case (b), the linearization in λ_0 does not exist. To avoid this difficulty we introduce the transformation

$$\lambda = \lambda(\mu) = \lambda_0 + s_0\mu^2, \tag{4.4}$$

$$s_0 := \operatorname{sgn}[D(\lambda_0)\cdot D'(\lambda_0)] \neq 0.$$

Note that s_0 is chosen in such a way that $\lambda(\mu) \notin \sigma$ for $\mu \neq 0$ and $|\mu| < \sqrt{\delta}$ (δ as in Lemma 4.1).

The transformation leads to the modified problem

$$-(pv')' + \tilde{q}v + f(x,v)v = \mu^2\tilde{s}v, \tag{4.5}$$

$$\begin{aligned}\tilde{\alpha}_{-1}(\mu)v(-N) - \tilde{\beta}_{-1}(\mu)v'(-N) &= 0,\\ \tilde{\alpha}_1(\mu)v(N) + \tilde{\beta}_1(\mu)v'(N) &= 0\end{aligned} \tag{4.6}$$

with μ as a new bifurcation parameter and coefficients

$$\begin{aligned}\tilde{q}(x) &:= q(x) - \lambda_0 s(x),\\ \tilde{s}(x) &:= s_0 s(x),\end{aligned}$$

$$\tilde{\alpha}_i(\mu) := \alpha_i(\lambda(\mu)), \quad \tilde{\beta}_i(\mu) := \beta_i(\lambda(\mu)) \quad (\mu \geq 0).$$

To extend $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ in a differentiable way for $\mu \leq 0$, we choose the appropriate sign in front of the square root and set

(i) in case $\varphi_2(\lambda_0) \neq 0$,

$$\begin{aligned}\tilde{\alpha}_i(\mu) &:= i\,[\varphi_1(\lambda(\mu)) - \varphi_2'(\lambda(\mu))] + \operatorname{sgn}[\mu D(\lambda(\mu))]\sqrt{D^2(\lambda(\mu)) - 4},\\ \tilde{\beta}_i(\mu) &:= 2\varphi_2(\lambda(\mu));\end{aligned} \tag{4.7}$$

(ii) in case $\varphi_2(\lambda_0) = 0$,

$$\begin{aligned}\tilde{\alpha}_i(\mu) &:= 2\varphi_1'(\lambda(\mu)),\\ \tilde{\beta}_i(\mu) &:= -i\,[\varphi_1(\lambda(\mu)) - \varphi_2'(\lambda(\mu))] + \operatorname{sgn}[\mu D(\lambda(\mu))]\sqrt{D^2(\lambda(\mu)) - 4}.\end{aligned} \tag{4.8}$$

Accordingly, we have to change the corresponding multiplier ϱ to

$$\tilde{\varrho}(\mu) := \frac{1}{2}\left\{D(\lambda(\mu)) - \operatorname{sgn}\left[\mu D(\lambda(\mu))\right]\sqrt{D^2(\lambda(\mu)) - 4}\right\}. \tag{4.9}$$

Note that $|\tilde{\varrho}(\mu)| > 1$ for $-\sqrt{\delta} < \mu < 0$ and $|\tilde{\varrho}(\mu)| < 1$ for $0 < \mu < \sqrt{\delta}$ by this choice.

Remark. Suppose (H0) and (H1) hold. If $v_\mu \not\equiv 0$ ($|\mu| < \sqrt{\delta}$) is a solution of (4.5), (4.6), then there exists a unique solution u_μ of (1.1) for $x \in \mathbb{R}$ such that $v_\mu = u_{\mu|[-N,N]}$. Since the functions

$$u_{-1}(\mu) := u_{\mu|(-\infty,-N)}, \qquad u_1(\mu) := u_{\mu|(N,\infty)}$$

satisfy

$$\begin{aligned} u_{-1}(x-2N,\mu) &= \tilde{\varrho}(\mu)u_{-1}(x,\mu) && (x < -N),\\ u_1(x+2N,\mu) &= \tilde{\varrho}(\mu)u_1(x,\mu) && (x > N), \end{aligned} \tag{4.10}$$

we can draw the following conclusions: if $\mu \le 0$, then $|\tilde{\varrho}(\mu)| \ge 1$ and $u_\mu \notin L^p(\mathbb{R})$ and if $\mu > 0$, then $|\tilde{\varrho}(\mu)| < 1$ and $u_\mu \in L^p(\mathbb{R})$.

We collect the required differentiability properties of the transformed coefficients:

LEMMA 4.2. (i) *$\tilde{\varrho}, \tilde{\alpha}_i, \tilde{\beta}_i$ are analytic* $(i = -1, 1)$.

(ii) $\tilde{\varrho}'(0) = -D(\lambda_0)\sqrt{|D'(\lambda_0)|}/2 \neq 0$.

(iii) (a) *If $\varphi_2(\lambda_0) \neq 0$, then*

$$\begin{aligned} &\tilde{\alpha}_i'(0) = -2\tilde{\varrho}'(0),\ \tilde{\beta}_i'(0) = 0 \quad (i = -1, 1),\\ &\tilde{\alpha}_{-1}(0) = -\tilde{\alpha}_1(0),\ \tilde{\beta}_{-1}(0) = \tilde{\beta}_1(0) = 2\varphi_2(\lambda_0) \neq 0. \end{aligned}$$

(b) *If $\varphi_2(\lambda_0) = 0$, then*

$$\begin{aligned} &\tilde{\alpha}_i'(0) = 0,\ \tilde{\beta}_i'(0) = -2\tilde{\varrho}'(0) \quad (i = -1, 1),\\ &\tilde{\alpha}_{-1}(0) = \tilde{\alpha}_1(0) = 2\varphi_1'(\lambda_0) \neq 0, \quad \tilde{\beta}_{-1}(0) = -\tilde{\beta}_1(0). \end{aligned}$$

Proof. $D(\lambda)$ is an analytic function of λ, hence $r(\lambda) := D^2(\lambda) - 4$ is analytic and $r(\lambda_0) = 0$, $r'(\lambda_0) = 2D(\lambda_0)D'(\lambda_0) \neq 0$. Hence there exists an analytic function d such that $r(\lambda) = (\lambda - \lambda_0)d(\lambda)$, $d(\lambda_0) = r'(\lambda_0) \neq 0$, and $d(\lambda) \neq 0$ ($\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta)\backslash\sigma$). The function $\sqrt{s_0 d(\lambda(\mu))}$ is analytic for $|\mu| < \sqrt{\delta}$ and

$$\begin{aligned} g(\mu) &= \operatorname{sgn}(\mu)\sqrt{D^2\left(\lambda(\mu)\right) - 4}\\ &= \operatorname{sgn}(\mu)\sqrt{s_0\mu^2 d\left(\lambda(\mu)\right)}\\ &= \mu\sqrt{s_0 d\left(\lambda(\mu)\right)} \end{aligned}$$

is analytic for $|\mu| < \sqrt{\delta}$,

$$\begin{aligned} g'(\mu) &= \sqrt{s_0 d\left(\lambda(\mu)\right)} + \mu s_0 d'(\lambda(\mu))\lambda'(\mu)/\left(2\sqrt{s_0 d(\lambda(\mu))}\right),\\ g'(0) &= \sqrt{s_0 d(\lambda_0)} = 2\sqrt{|D'(\lambda_0)|}. \end{aligned}$$

Since $D(\lambda(\mu))$, $\varphi_1(\lambda(\mu))$, $\varphi_2(\lambda(\mu))$ are analytic and $D(\lambda(\mu)) \neq 0$ ($|\mu| < \sqrt{\delta}$), $\tilde{\varrho}, \tilde{\alpha}_i, \tilde{\beta}_i$ are analytic and the formulae in (ii) and (iii) are straightforward. □

We now solve equation (4.5), (4.6) by a shooting method. Let $v(x) = v(x, \varepsilon, \mu)$ denote the maximal solution of equation (4.5) satisfying the initial conditions

$$\begin{aligned} v(-N) &= \varepsilon\tilde{\beta}_{-1}(\mu),\\ v'(-N) &= \varepsilon\tilde{\alpha}_{-1}(\mu), \end{aligned} \tag{4.11}$$

and denote by $I(\varepsilon,\mu)$ the maximal interval of existence of v; by this choice v fulfills the boundary condition (4.6) at $-N$. We will show that for small ε there is a $\mu = M(\varepsilon)$ such that $v(x,\varepsilon,M(\varepsilon))$ exists on $[-N,N]$ and satisfies the boundary condition at N; furthermore, close to the bifurcation point, the dominant part of the bifurcating solution is given (as usual) by ψ, the solution of the linearized equation.

THEOREM 4.3. *If* (H0), (H1), *and* (H2) *hold there exists a (local) branch of solutions* $(v_\varepsilon, M(\varepsilon))$ *of* (4.5), (4.6) *bifurcating at* $(0,0)$. *More precisely, there exist* $\varepsilon_0 > 0$, $\mu_0 \in (0,\sqrt{\delta})$, *and a continuous function* $M : [-\varepsilon_0,\varepsilon_0] \to [-\mu_0,\mu_0]$ *such that*

(i) $M(0) = 0$;

(ii) $[-N,N] \subseteq I(\varepsilon, M(\varepsilon))$;

(iii) $v_\varepsilon(x) = v(x,\varepsilon,M(\varepsilon))$ *is a nontrivial solution of* (4.5), (4.6) *for* $\mu = M(\varepsilon)$, $\varepsilon \neq 0$;

(iv) $v_\varepsilon(x) = \varepsilon(\psi + \gamma(\varepsilon))(x)$ *where* ψ *is defined as in* (4.3) *and* $\gamma : [-\varepsilon_0,\varepsilon_0] \to (C^1[-N,N], \| \ \|)$ *is a continuous function satisfying* $\|\gamma(0)\| = 0$ *for* $\|y\| := \max\{\|y\|_\infty, \|y'\|_\infty\}$;

(v) *any solution* (v,μ) *in a neighbourhood of* $(0,0)$ *is of the form* $\mu = M(\varepsilon)$, $v = v_\varepsilon$;

(vi) *there exists a continuous function* $c : [-\varepsilon_0,\varepsilon_0] \to \mathbb{R}$ *with* $c(0) = 0$ *such that in case* $\varphi_2(\lambda_0) \neq 0$,

$$\begin{aligned} v_\varepsilon(N) &= j\varepsilon(1 + c(\varepsilon))\tilde{\beta}_1(M(\varepsilon)), \\ v_\varepsilon'(N) &= -j\varepsilon(1 + c(\varepsilon))\tilde{\alpha}_1(M(\varepsilon)), \end{aligned} \tag{4.12}$$

and in case $\varphi_2(\lambda_0) = 0$,

$$\begin{aligned} v_\varepsilon(N) &= -j\varepsilon(1 + c(\varepsilon))\tilde{\beta}_1(M(\varepsilon)), \\ v_\varepsilon'(N) &= j\varepsilon(1 + c(\varepsilon))\tilde{\alpha}_1(M(\varepsilon)) \end{aligned} \tag{4.13}$$

with $j = D(\lambda_0)/2$.

Proof. To get global existence of the solution $v(x,\varepsilon,\mu)$ of the initial value problem (4.5), (4.11), we modify the nonlinearity outside the strip $\mathbb{R} \times [-y_0,y_0]$:

$$\tilde{f}(x,y) := \begin{cases} f(x,y), & |y| \le y_0, \\ f(x,y_0), & y > y_0, \\ f(x,-y_0), & y < -y_0. \end{cases}$$

Then $\tilde{f}(x,y)\cdot y$ is continuous with respect to y and satisfies a global Lipschitz condition. It is sufficient to prove Theorem 4.3 for equation (4.5) with f replaced by $\tilde{f}$ since the continuous dependence of v_ε on ε implies that $\|v_\varepsilon\|_\infty \le y_0$ $(|\varepsilon| \le \varepsilon_0)$ for ε_0 sufficiently small.

Let $v = v(\cdot,\varepsilon,\mu)$ be the solution of the initial value problem (4.5), (4.11) with f replaced by $\tilde{f}$. Then v has the following properties:

(i) $I(\varepsilon,\mu) = \mathbb{R}$ for all $\varepsilon \in \mathbb{R}$, $|\mu| < \sqrt{\delta}$.

(ii) $v(x,0,\mu) \equiv 0$ for all $x \in \mathbb{R}$, $|\mu| < \sqrt{\delta}$.

(iii) We define $V : \mathbb{R} \times \mathbb{R} \times (-\sqrt{\delta},\sqrt{\delta}) \to \mathbb{R}$ as the solution of the initial value problem

$$-(pv')' + \tilde{q}v + \tilde{f}(x,\varepsilon v)v = \mu^2 \tilde{s}v \tag{4.14}$$

with initial conditions

$$v(-N) = \tilde{\beta}_{-1}(\mu), \qquad v'(-N) = \tilde{\alpha}_{-1}(\mu). \tag{4.15}$$

Note that $V(x,\varepsilon,\mu) = v(x,\varepsilon,\mu)/\varepsilon$ for $\varepsilon \neq 0$ and $V(x,0,0) = \psi(x)$.

(iv) $V(x,\varepsilon,\mu)$ is continuously differentiable with respect to μ and $W(x,\varepsilon,\mu) := \frac{\partial V}{\partial \mu}(x,\varepsilon,\mu)$ solves the initial value problem

$$\begin{aligned} -(pw')' + \tilde{q}w + \left(\tfrac{\partial \tilde{f}}{\partial y}(x,\varepsilon V(x,\varepsilon,\mu))\varepsilon V(x,\varepsilon,\mu) + \tilde{f}(x,\varepsilon V(x,\varepsilon,\mu))\right) w \\ = \mu^2 \tilde{s} w + 2\mu \tilde{s} V(x,\varepsilon,\mu), \end{aligned} \tag{4.16}$$

$$w(-N) = \tilde{\beta}'_{-1}(\mu), \quad w'(-N) = \tilde{\alpha}'_{-1}(\mu). \tag{4.17}$$

In particular, for $\varepsilon = \mu = 0$ we get

$$W(x,0,0) = \tilde{\beta}'_{-1}(0)\varphi_1(x,\lambda_0) + \tilde{\alpha}'_{-1}(0)\varphi_2(x,\lambda_0).$$

Now we can set up the defining condition for the shooting procedure which will be solved by the implicit function theorem. Define $g : \mathbb{R} \times (-\sqrt{\delta}, \sqrt{\delta}) \to \mathbb{R}$ by

$$g(\varepsilon,\mu) := \tilde{\alpha}_1(\mu)V(N,\varepsilon,\mu) + \tilde{\beta}_1(\mu)V'(N,\varepsilon,\mu).$$

If (v,μ) is a nontrivial solution of the boundary value problem (4.5), (4.6), then there exists $\varepsilon \neq 0$ (namely, $\varepsilon = v(-N)/\tilde{\beta}_{-1}(\mu)$ if $\varphi_2(\lambda_0) \neq 0$ and $\varepsilon = v'(-N)/\tilde{\alpha}_{-1}(\mu)$ if $\varphi_2(\lambda_0) = 0$) such that $v(x) = \varepsilon V(x,\varepsilon,\mu)$ and

$$\begin{aligned} \varepsilon g(\varepsilon,\mu) &= \tilde{\alpha}_1(\mu)\varepsilon V(N,\varepsilon,\mu) + \tilde{\beta}_1(\mu)\varepsilon V'(N,\varepsilon,\mu) \\ &= \tilde{\alpha}_1(\mu)v(N,\varepsilon,\mu) + \tilde{\beta}_1(\mu)v'(N,\varepsilon,\mu) \\ &= \tilde{\alpha}_1(\mu)v(N) + \tilde{\beta}_1(\mu)v'(N) \\ &= 0. \end{aligned}$$

On the other hand, if $g(\varepsilon,\mu) = 0$ for some $\varepsilon \neq 0$, $|\mu| < \sqrt{\delta}$, then by construction $v(x,\varepsilon,\mu)$ is a nontrivial solution of (4.5), (4.6).

Then by Lemma 4.1,

$$\begin{aligned} g(0,0) &= \tilde{\alpha}_1(0)V(N,0,0) + \tilde{\beta}_1(0)V'(N,0,0) \\ &= \tilde{\alpha}_1(0)\psi(N) + \tilde{\beta}_1(0)\psi'(N) = 0. \end{aligned}$$

The function g is continuous with respect to ε, and by Lemma 4.2 and property (iv) it is continuously differentiable with respect to μ.

$$\frac{dg}{d\mu}(\varepsilon,\mu) = \tilde{\alpha}'_1(\mu)V(N,\varepsilon,\mu) + \tilde{\beta}'_1(\mu)V'(N,\varepsilon,\mu) + \tilde{\alpha}_1(\mu)W(N,\varepsilon,\mu) + \tilde{\beta}_1(\mu)W'(N,\varepsilon,\mu),$$

hence

$$\begin{aligned} \frac{dg}{d\mu}(0,0) = {} & \tilde{\alpha}'_1(0)\left[\tilde{\beta}_{-1}(0)\varphi_1(\lambda_0) + \tilde{\alpha}_{-1}(0)\varphi_2(\lambda_0)\right] \\ & + \tilde{\beta}'_1(0)\left[\tilde{\beta}_{-1}(0)\varphi'_1(\lambda_0) + \tilde{\alpha}_{-1}(0)\varphi'_2(\lambda_0)\right] \\ & + \tilde{\alpha}_1(0)\left[\tilde{\beta}'_{-1}(0)\varphi_1(\lambda_0) + \tilde{\alpha}'_{-1}(0)\varphi_2(\lambda_0)\right] \\ & + \tilde{\beta}_1(0)\left[\tilde{\beta}'_{-1}(0)\varphi'_1(\lambda_0) + \tilde{\alpha}'_{-1}(0)\varphi'_2(\lambda_0)\right]. \end{aligned}$$

In case $\varphi_2(\lambda_0) \neq 0$, by Lemma 4.2 we obtain

$$\begin{aligned} \frac{dg}{d\mu}(0,0) &= -2\tilde{\varrho}'(0)\left[2\varphi_2(\lambda_0)\varphi_1(\lambda_0) + \tilde{\alpha}_{-1}(0)\varphi_2(\lambda_0)\right] \\ &\quad + \tilde{\alpha}_1(0)\left[-2\tilde{\varrho}'(0)\varphi_2(\lambda_0)\right] + 2\varphi_2(\lambda_0)\left[-2\tilde{\varrho}'(0)\varphi'_2(\lambda_0)\right] \\ &= -4\tilde{\varrho}'(0)\varphi_2(\lambda_0)\left[\varphi_1(\lambda_0) + \varphi'_2(\lambda_0)\right] \\ &= +8\sqrt{|D'(\lambda_0)|}\varphi_2(\lambda_0) \neq 0. \end{aligned}$$

Similarly, we obtain $\frac{dg}{d\mu}(0,0) = 8\sqrt{|D'(\lambda_0)|}\varphi_1'(\lambda_0) \neq 0$ in case $\varphi_2(\lambda_0) = 0$.

Applying the implicit function theorem, we obtain that there is $\varepsilon_0 > 0$ and a continuous function $M : [-\varepsilon_0, \varepsilon_0] \to [-\mu_0, \mu_0]$ such that for each $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$ there is a unique $\mu = M(\varepsilon)$ such that $g(\varepsilon, M(\varepsilon)) \equiv 0$ $(-\varepsilon_0 \leq \varepsilon \leq \varepsilon_0)$; in addition, $M(0) = 0$. By construction, $v_\varepsilon := v(x, \varepsilon, M(\varepsilon))$ is a solution of the boundary value problem (4.5), (4.6).

To prove (iv) define γ by

$$\gamma(\varepsilon)(x) := V(x, \varepsilon, M(\varepsilon)) - \psi(x).$$

Clearly, γ is continuous and because of $V(x, 0, 0) = \psi(x)$ we obtain

$$\|\gamma(0)\| = 0,$$

and by definition of γ,

$$v_\varepsilon(x) = v(x, \varepsilon, M(\varepsilon)) = \varepsilon V(x, \varepsilon, M(\varepsilon)) = \varepsilon(\psi + \gamma(\varepsilon))(x).$$

To prove (v) assume that (v, μ) is a nontrivial solution of (4.5), (4.6) such that $|\mu| \leq \mu_0$ and $\max\{\|v\|_\infty, \|v'\|_\infty\} \leq \hat{\varepsilon} = 2\varepsilon_0 \min_{|\mu| \leq \mu_0} \{|\varphi_2(\lambda(\mu))|\}$ in case $\varphi_2(\lambda_0) \neq 0$. Then $|v(-N)| \leq \hat{\varepsilon}$ and $\varepsilon := v(-N)/(2\varphi_2(\lambda(\mu)))$ satisfies $|\varepsilon| \leq \varepsilon_0$; moreover, $v(-N) = 2\varepsilon\varphi_2(\lambda(\mu)) = \varepsilon\tilde{\beta}_{-1}(\mu)$. Since v satisfies the boundary condition at $-N$, we obtain $v'(-N) = \varepsilon\tilde{\alpha}_{-1}(\mu)$. By definition of v, $v = v(\cdot, \varepsilon, \mu)$ and $g(\varepsilon, \mu) = 0$, hence $\mu = M(\varepsilon)$ by the uniqueness of μ. The case $\varphi_2(\lambda_0) = 0$ is treated similarly with the choice $\varepsilon = v'(-N)/(2\varphi_1'(\lambda(\mu)))$, $\hat{\varepsilon} = 2\varepsilon_0 \min_{|\mu| \leq \mu_0} \{|\varphi_1'(\lambda(\mu))|\}$.

Since the solution v_ε satisfies the boundary condition at N, there exists a continuous function $\tau : [-\varepsilon_0, \varepsilon_0] \to \mathbb{R}$ such that for all $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$,

$$v_\varepsilon(N) = \tau(\varepsilon)\tilde{\beta}_1(M(\varepsilon)), \qquad v_\varepsilon'(N) = -\tau(\varepsilon)\tilde{\alpha}_1(M(\varepsilon)).$$

In case $\varphi_2(\lambda_0) \neq 0$ we define the continuous function $c : [-\varepsilon_0, \varepsilon_0] \to \mathbb{R}$ by

$$c(\varepsilon) := \left[j\gamma(\varepsilon)(N) - \int_0^{M(\varepsilon)} \tilde{\beta}_1'(t)dt \right] / \tilde{\beta}_1(M(\varepsilon)),$$

where $j = D(\lambda_0)/2$ $(j = \pm 1)$, hence $c(0) = j\gamma(0)(N)/\tilde{\beta}_1(0) = 0$ and, by Lemmas 4.1 and 4.2,

$$\begin{aligned}
v_\varepsilon(N) &= \varepsilon\left(\psi + \gamma(\varepsilon)\right)(N) \\
&= \varepsilon\left(j\psi(-N) + \gamma(\varepsilon)(N)\right) \\
&= \varepsilon\left(j\tilde{\beta}_{-1}(0) + \gamma(\varepsilon)(N)\right) \\
&= \varepsilon\left(j\tilde{\beta}_1(0) + \gamma(\varepsilon)(N)\right) \\
&= \varepsilon\left(j\tilde{\beta}_1(M(\varepsilon)) - j\int_0^{M(\varepsilon)} \tilde{\beta}'(t)dt + \gamma(\varepsilon)(N)\right) \\
&= \varepsilon\tilde{\beta}_1\left(M(\varepsilon)\right) j(1 + c(\varepsilon)).
\end{aligned}$$

Consequently, $\tau(\varepsilon) = (1 + c(\varepsilon))\varepsilon j$ and $v_\varepsilon'(N) = -\varepsilon\tilde{\alpha}_1(M(\varepsilon))(1 + c(\varepsilon))j$. The assertion in the case $\varphi_2'(\lambda_0) = 0$ follows in the same way. □

Theorem 4.3 states that λ_0 (i.e., $\mu = 0$) is always a bifurcation point for the reduced problem. The function $\mu = M(\varepsilon)$ determines the kind and the direction of bifurcation. We will see later that there are important consequences for the original problem since the multiplier $\tilde{\varrho}(\mu)$ is contracting (resp., expanding) if $\mu > 0$ (resp., $\mu < 0$).

THEOREM 4.4. *Suppose that* (H0), (H1), *and* (H2) *hold and choose* ε_0, μ_0, M, *and* γ *as in Theorem* 4.3 *and* ψ *as in* (4.3).

Then there exists $\varepsilon_1 \in \mathbb{R}$, $0 < \varepsilon_1 \leq \varepsilon_0$, *and a continuous function* $R : [-\varepsilon_1, \varepsilon_1] \to \mathbb{R}$ *with* $R(0) = 0$ *such that, for* $|\varepsilon| \leq \varepsilon_1$,

$$M(\varepsilon) = \frac{1}{T} \int_{-N}^{N} f(x, \varepsilon(\psi + \gamma(\varepsilon))(x))(\psi + \gamma(\varepsilon))(x)\psi(x)dx(1 + R(\varepsilon)), \tag{4.18}$$

where

$$T = \begin{cases} -4p(N)D(\lambda_0)\sqrt{|D'(\lambda_0)|}\varphi_2(\lambda_0) & \text{if} \quad \varphi_2(\lambda_0) \neq 0, \\ 4p(N)D(\lambda_0)\sqrt{|D'(\lambda_0)|}\varphi_1'(\lambda_0) & \text{if} \quad \varphi_2(\lambda_0) = 0. \end{cases}$$

Remark. If $\lambda_0 \in \partial_r\sigma$ then $T > 0$; if $\lambda_0 \in \partial_\ell\sigma$ then $T < 0$. Here $\partial_r\sigma$ (resp., $\partial_\ell\sigma$) denotes the set of right (resp., left) end points of σ.

Proof. For $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$ let $(v_\varepsilon, M(\varepsilon))$ be the solution of (4.5), (4.6) given by Theorem 4.3. Multiplying equation (4.5) with ψ and integrating over $[-N, N]$ gives

$$\begin{aligned} -\int_{-N}^{N} (pv_\varepsilon')'\psi(x)dx + \int_{-N}^{N} \tilde{q}v_\varepsilon\psi(x)dx + \int_{-N}^{N} f(x, v_\varepsilon(x))v_\varepsilon(x)\psi(x)dx \\ = M^2(\varepsilon) \int_{-N}^{N} \tilde{s}(x)v_\varepsilon(x)\psi(x)dx. \end{aligned}$$

Using the fact that ψ is a solution of the linear equation for $\mu = 0$, after integration by parts we get

$$B(\varepsilon) + \int_{-N}^{N} f(x, v_\varepsilon(x))v_\varepsilon(x)\psi(x)dx = M^2(\varepsilon) \int_{-N}^{N} \tilde{s}(x)v_\varepsilon(x)\psi(x)dx, \tag{4.19}$$

where the boundary terms are collected in

$$\begin{aligned} B(\varepsilon) = &-p(N)\left[v_\varepsilon'(N)\psi(N) - v_\varepsilon(N)\psi'(N)\right] \\ &+p(-N)\left[v_\varepsilon'(-N)\psi(-N) - v_\varepsilon(-N)\psi'(-N)\right]. \end{aligned}$$

By (4.12) and the $2N$-periodicity of p,

$$\begin{aligned} B(\varepsilon) = \varepsilon p(N) \Big[& js_B(1 + c(\varepsilon)) \left(\tilde{\alpha}_1(M(\varepsilon))\psi(N) + \tilde{\beta}_1(M(\varepsilon))\psi'(N)\right) \\ &+ \left(\tilde{\alpha}_{-1}(M(\varepsilon))\psi(-N) - \tilde{\beta}_{-1}(M(\varepsilon))\psi'(-N)\right)\Big] \end{aligned}$$

with $j = \operatorname{sgn} D(\lambda_0)$ and $s_B = 1$ if $\varphi_2(\lambda_0) \neq 0$, and $s_B = -1$ if $\varphi_2(\lambda_0) = 0$. Using the Taylor expansions

$$\tilde{\alpha}_i(M(\varepsilon)) = \tilde{\alpha}_i(0) + \tilde{\alpha}_i'(0)M(\varepsilon) + \int_0^{M(\varepsilon)} (M(\varepsilon) - t)\tilde{\alpha}_i''(t)dt,$$

$$\tilde{\beta}_i(M(\varepsilon)) = \tilde{\beta}_i(0) + \tilde{\beta}_i'(0)M(\varepsilon) + \int_0^{M(\varepsilon)} (M(\varepsilon) - t)\tilde{\beta}_i''(t)dt$$

and the equation $\tilde{\alpha}_i(0)\psi(iN) + i\tilde{\beta}_i(0)\psi'(iN) = 0$, we obtain

$$B(\varepsilon) = \varepsilon p(N)\left\{js_B\left[\tilde{\alpha}_1'(0)\psi(N) + \tilde{\beta}_1'(0)\psi'(N)\right] + \left[\tilde{\alpha}_{-1}'(0)\psi(-N) - \tilde{\beta}_{-1}'(0)\psi'(-N)\right]\right\} M(\varepsilon) + \varepsilon M(\varepsilon)C(\varepsilon),$$

where $C : [-\varepsilon_0, \varepsilon_0] \to \mathbb{R}$ is the continuous function satisfying $C(0) = 0$ given by

$$\begin{aligned} C(\varepsilon) &= p(N)js_B c(\varepsilon)[\tilde{\alpha}_1'(0)\psi(N) + \tilde{\beta}_1'(0)\psi'(N)] \\ &\quad + p(N)js_B \frac{(1 + c(\varepsilon))}{M(\varepsilon)} \left[\int_0^{M(\varepsilon)} (M(\varepsilon) - t)\tilde{\alpha}_1''(t)dt\psi(N) \right. \\ &\qquad \left. + \int_0^{M(\varepsilon)} (M(\varepsilon) - t)\tilde{\beta}_1''(t)dt\psi'(N)\right] \\ &\quad + \frac{p(N)}{M(\varepsilon)}\left[\int_0^{M(\varepsilon)} (M(\varepsilon) - t)\tilde{\alpha}_{-1}''(t)\psi(-N) - \int_0^{M(\varepsilon)} (M(\varepsilon) - t)\tilde{\beta}_{-1}''(t)dt\psi'(-N)\right]. \end{aligned}$$

Hence, by Lemmas 4.1 and 4.2,

$$B(\varepsilon) = -T\varepsilon M(\varepsilon) + \varepsilon M(\varepsilon)C(\varepsilon).$$

Dividing (4.19) by ε gives

$$TM(\varepsilon) - E(\varepsilon)M(\varepsilon) = \int_{-N}^{N} f(x, v_\varepsilon(x))\ (\psi + \gamma(\varepsilon))(x)\psi(x)dx,$$

where

$$\begin{aligned} &E : [-\varepsilon_0, \varepsilon_0] \to \mathbb{R}, \\ &E(\varepsilon) := C(\varepsilon) - M(\varepsilon)\int_{-N}^{N} \tilde{s}(x)(\psi(x) + \gamma(\varepsilon)(x))\psi(x)dx \end{aligned}$$

is a continuous function with $E(0) = 0$.

Choose $\varepsilon_1 \in (0, \varepsilon_0)$ such that $E(\varepsilon) \neq T$ $(|\varepsilon| \leq \varepsilon_1)$; then

$$M(\varepsilon) = \frac{1}{T}(1 + R(\varepsilon))\int_{-N}^{N} f(x, v_\varepsilon(x))(\psi + \gamma(\varepsilon))\psi(x)dx,$$

where $R : [-\varepsilon_1, \varepsilon_1] \to \mathbb{R}$, $R(\varepsilon) := E(\varepsilon)/(T - E(\varepsilon))$ is continuous and satisfies $R(0) = 0$. □

For general f it is not yet possible to work out the dominant part of $M(\varepsilon)$ in terms of ε by (4.18) since the right-hand side depends on $M(\varepsilon)$ in form of $\gamma(\varepsilon)$.

To evaluate formula (4.18) determining $M(\varepsilon)$ explicitly, we need more specific assumptions on the nonlinearity f. As a typical application we treat the case

(H3) $f(x,y) = g(x,y) + h(x,y)$,
$g(x,y) = \begin{cases} r_1(x)|y|^\sigma & (y \geq 0), \\ r_{-1}(x)|y|^\sigma & (y < 0), \end{cases}$
where $r_j : [-N, N] \to \mathbb{R}$ is piecewise continuous, $r_1 \not\equiv 0$ or $r_{-1} \not\equiv 0$, and $\sigma > 0$.
Furthermore, there exists $\omega > 0, \gamma > 0$ such that

$$|h(x,y)| \leq \omega |y|^{\sigma+\gamma} \quad (|y| \leq y_0).$$

For the function ψ in (4.3) we use the decomposition $\psi(x) = \psi_1(x) - \psi_{-1}(x)$ into a positive (ψ_1) and negative (ψ_{-1}) part, and we define ($j = -1, 1$)

$$T_j := \int_{-N}^{N} r_j(x)|\psi_j(x)|^{\sigma+2} dx,$$

$$S_j := \int_{-N}^{N} r_j(x)|\psi_{-j}(x)|^{\sigma+2} dx.$$

COROLLARY 4.5. *Suppose* (H0), (H1), (H2), *and* (H3) *hold and let* ε_1, M, T *be as in Theorem* 4.4.

(i) *If* $T_1 + T_{-1} \neq 0$ *then there exists a continuous function* $R_+ : [0, \varepsilon_1] \to \mathbb{R}$ *with* $R_+(0) = 0$ *such that*

$$M(\varepsilon) = \frac{T_1 + T_{-1}}{T} |\varepsilon|^\sigma (1 + R_+(\varepsilon)) \quad (0 \leq \varepsilon \leq \varepsilon_1).$$

(ii) *If* $S_1 + S_{-1} \neq 0$ *then there exists a continuous function* $R_- : [-\varepsilon_1, 0] \to \mathbb{R}$ *with* $R_-(0) = 0$ *such that*

$$M(\varepsilon) = \frac{S_1 + S_{-1}}{T} |\varepsilon|^\sigma (1 + R_-(\varepsilon)) \quad (-\varepsilon_1 \leq \varepsilon \leq 0).$$

Remark. Since solutions of the original problem only exist if $M(\varepsilon) > 0$, we mention two special situations where $M(\varepsilon) > 0$ is guaranteed by Corollary 4.5.

Example 4.6. $r_1 = r_{-1}$ (i.e., $g(x,y)y$ odd). Then $M(\varepsilon) > 0$ for $\varepsilon \neq 0$ if
(a) $\lambda_0 \in \partial_r \sigma$ and $r_1 \geq 0$ or
(b) $\lambda_0 \in \partial_\ell \sigma$ and $r_1 \leq 0$.

Example 4.7. $r_1 = -r_{-1}$ (i.e., $g(x,y)y$ even). Then $T_1 = -S_{-1}$ and $T_{-1} = -S_1$ and

$$\begin{aligned} T_1 + T_{-1} &= \int_{-N}^{N} r_1(x)|\psi_1(x)|^{\sigma+2} dx - \int_{-N}^{N} r_1(x)|\psi_{-1}(x)|^{\sigma+2} dx \\ &= \int_{-N}^{N} r_1(x) \left[|\psi_1(x)|^{\sigma+2} - |\psi_{-1}(x)|^{\sigma+2} \right] dx \\ &= \int_{-N}^{N} r_1(x) \left[|\psi_1(x)|^{\sigma+1}\psi_1(x) - |\psi_{-1}(x)|^{\sigma+1}\psi_{-1}(x) \right] dx \\ &= \int_{-N}^{N} r_1(x)|\psi(x)|^{\sigma+1}\psi(x) dx. \end{aligned}$$

If $T_1 + T_{-1} \neq 0$ then either $T_1 + T_{-1} > 0$ or $S_1 + S_{-1} = -(T_1 + T_{-1}) > 0$; hence $M(\varepsilon) > 0$ either for $\varepsilon > 0$ or $\varepsilon < 0$.

Corollary 4.5 is a direct consequence of the following estimates collected in the following lemma.

LEMMA 4.8. *Suppose* (H0), (H1), (H2), *and* (H3) *hold.*

There exists $\Omega > 0$ *and a continuous function* $C : [-\varepsilon_0, \varepsilon_0] \to \mathbb{R}$ *with* $C(0) = 0$ *such that, for* $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$,

(i)

$$\left| \int_{-N}^{N} [f(x, \varepsilon(\psi + \gamma(\varepsilon))(x)) - g(x, \varepsilon(\psi + \gamma(\varepsilon))(x))] [\psi + \gamma(\varepsilon)](x)\psi(x)dx \right| \leq \Omega|\varepsilon|^{\sigma+\gamma},$$

(ii)

$$\left| \int_{-N}^{N} [g(x, \varepsilon(\psi(x) + \gamma(\varepsilon)(x)))(\psi(x) + \gamma(\varepsilon)(x)) - g(x, \varepsilon\psi(x))\psi(x)] \, dx \right| = C(\varepsilon)|\varepsilon|^{\sigma},$$

(iii)

$$\int_{-N}^{N} g(x, \varepsilon\psi(x))\psi^2(x)dx = \begin{cases} (T_1 + T_{-1})|\varepsilon|^{\sigma} & (\varepsilon > 0), \\ (S_1 + S_{-1})|\varepsilon|^{\sigma} & (\varepsilon < 0). \end{cases}$$

5. Conclusions for the original problem. Theorem 4.3 guarantees that every $\lambda_0 \in \partial\sigma$ is a bifurcation point for the reduced problem (4.5), (4.6) and that there is a local branch $(v_\varepsilon, M(\varepsilon))$. The direction of bifurcation is determined by the sign of $\mu = M(\varepsilon)$; moreover, if $\mu \leq 0$ then $|\tilde{\varrho}(\mu)| \geq 1$ and the solution of the reduced problem does not lead to a solution of the original problem since $\tilde{\varrho}(\mu)$ acts as an expanding multiplier. For $\mu > 0$, however, solutions in $L^p(\mathbb{R})$ can be constructed out of solutions of the reduced problem. Precisely, if $(v_\varepsilon, M(\varepsilon))$ is a solution of (4.5), (4.6) and $M(\varepsilon) > 0$, we can extend v_ε in the spirit of (3.7) to a solution $u_\lambda \in L^p(\mathbb{R})$ of (1.1) by using

$$\lambda = \lambda_0 + s_0 M^2(\varepsilon) \tag{5.1}$$

and

$$u_\lambda(x) = \begin{cases} \varepsilon u_{-1}(x, \lambda) & (x \leq -N), \\ v_\varepsilon(x) & (-N \leq x \leq N), \\ \kappa\varepsilon(1 + c(\varepsilon))u_1(x - 2N, \lambda) & (x \geq N), \end{cases} \tag{5.2}$$

where $\kappa \in \{-1, 1\}$ must be chosen appropriately since the coefficients $c_{-1}(\lambda), c_1(\lambda)$ turn out to be (in case $\varphi_2(\lambda_0) \neq 0$)

$$\begin{aligned} c_{-1}(\lambda) &= v_\varepsilon(-N)/u_{-1}(-N, \lambda) = \varepsilon\tilde{\beta}_{-1}(M(\varepsilon))/\tilde{\beta}_{-1}(M(\varepsilon)) = \varepsilon, \\ c_1(\lambda) &= v_\varepsilon(N)/u_1(-N, \lambda) = \kappa\varepsilon(1 + c(\varepsilon))\tilde{\beta}_1(M(\varepsilon))/\tilde{\beta}_1(M(\varepsilon)) = \varepsilon(1 + c(\varepsilon)), \end{aligned}$$

where c is given in Theorem 4.3 (vi).

The main results of this paper are collected in Theorems 5.1 and 5.2. Although they involve the existence of the solutions v_ε of the reduced problem, we do not need these explicitly. Their existence is guaranteed by §4; the only quantity needed is

the direction of the bifurcation given by the sign of $M(\varepsilon)$. In many cases, such as under the hypothesis of Corollary 4.5, this can be worked out a priori by using the eigenfunction of the linear problem; in those cases $\operatorname{sgn} M(\varepsilon)$ can be read off from the expressions in Corollary 4.5, for example.

We first treat the case $M(\varepsilon) \leq 0$ ($|\varepsilon| \leq \varepsilon_0$) and show that there is no bifurcation from 0 at λ_0 for the original problem. If solutions $u_\lambda \in L^p(\mathbb{R})$ nevertheless exist for $|\lambda_0 - \lambda| \leq \mu_0^2$ and $\lambda \notin \sigma$, then asymptotic bifurcation at λ_0 is the only possible situation.

THEOREM 5.1. *Suppose* (H0), (H1), *and* (H2) *hold and let* ε_0, μ_0, *and* $(v_\varepsilon, M(\varepsilon))$ *be defined as in Theorem* 4.3. *Assume, in addition, that* $M(\varepsilon) \leq 0$ ($|\varepsilon| \leq \bar{\varepsilon_0}$) *for some* $\bar{\varepsilon_0} \in (0, \varepsilon_0]$. *Then every nontrivial solution* $u_\lambda \in L^p(\mathbb{R})$ *of* (1.1) *with* $\lambda \notin \sigma$, $|\lambda - \lambda_0| \leq \mu_0^2$ *satisfies the estimate*

$$\|u_\lambda\|_p^p \geq \frac{C}{\sqrt{|\lambda - \lambda_0|}}, \tag{5.3}$$

where C *is a positive constant independent of* λ, u_λ. *In particular, there is no bifurcation from* 0 *at* λ_0 *for* (1.1).

Proof. If (u_λ, λ) is a nontrivial solution of (1.1) then, by Theorem 3.1, (v_λ, λ) with $v_\lambda := u_{\lambda|[-N,N]}$ is a nontrivial solution of (3.5), (3.6). Defining $\mu = +\sqrt{|\lambda - \lambda_0|}$, we also obtain a nontrivial solution of (4.5), (4.6) satisfying $g(\varepsilon, \mu) = 0$, where $\varepsilon = u_\lambda(-N)/\tilde{\beta}_{-1}(\mu)$ in case $\varphi_2(\lambda_0) \neq 0$ or $\varepsilon = u_\lambda'(-N)/\tilde{\alpha}_{-1}(\mu)$ in case $\varphi_2(\lambda_0) = 0$. If $|\varepsilon| \leq \varepsilon_0$ then, by uniqueness of μ through the implicit function theorem, $0 < \mu = M(\varepsilon)$; then $|\varepsilon| > \bar{\varepsilon_0}$ since $M(\varepsilon) \leq 0 (|\varepsilon| \leq \bar{\varepsilon_0})$; hence $|\varepsilon| \geq \bar{\varepsilon_0}$ always holds. Furthermore, $|\tilde{\varrho}(\mu)| < 1$ and by Lemma 4.2 there exists a continuous function $\vartheta : [-\mu_0, \mu_0] \to \mathbb{R}$ such that

$$|\tilde{\varrho}(\mu)|^p = |\tilde{\varrho}(0)|^p + p|\tilde{\varrho}(0)|^{p-1} \operatorname{sgn}(\tilde{\varrho}(0)) \tilde{\varrho}'(0)\mu + \vartheta(\mu)\mu^2.$$

Hence, using $\operatorname{sgn}(\tilde{\varrho}(0)) D(\lambda_0) = 2$,

$$\begin{aligned} 1 - |\tilde{\varrho}(\mu)|^p &= -p \operatorname{sgn}(\tilde{\varrho}(0)) \tilde{\varrho}'(0)\mu - \vartheta(\mu)\mu^2 \\ &= +\frac{p}{2} \operatorname{sgn}(\tilde{\varrho}(0)) D(\lambda_0) \sqrt{|D'(\lambda_0)|}\mu - \vartheta(\mu)\mu^2, \end{aligned}$$

$$\frac{1}{1 - |\tilde{\varrho}(\mu)|^p} \geq \frac{1}{p\sqrt{|D'(\lambda_0)|} + \vartheta_0 \mu} \frac{1}{\mu} \geq \tilde{C}/\mu,$$

where $\vartheta_0 := \max_{0 \leq \mu \leq \mu_0} |\vartheta(\mu)|$ and $\tilde{C}$ is some positive constant.

The assertion of the theorem now follows by the relation of the norms in Lemma 3.2:

$$\begin{aligned} \|u_\lambda\|_p^p = \int_{-N}^{N} |v_\lambda(x)|^p dx + \frac{1}{1 - |\tilde{\varrho}(\mu)|^p} \Bigg[&|c_{-1}(\lambda)\varrho(\lambda)|^p \int_{-N}^{N} |u_{-1}(x, \lambda)|^p dx \\ &+ |c_1(\lambda)|^p \int_{-N}^{N} |u_1(x, \lambda)|^p dx \Bigg] \\ \geq |c_{-1}(\lambda)|^p \tilde{C} \eta \frac{1}{\mu}, \end{aligned}$$

where $\eta = \min_{|\mu|\le\mu_0} \int_{-N}^{N} |u_{-1}(x,\lambda(\mu))|^p dx |\tilde{\varrho}(\mu)|^p$.

Since $c_{-1}(\lambda) = \varepsilon$ and $|\varepsilon| \ge \bar{\varepsilon_0}$ because of $M(\varepsilon) \le 0$ ($|\varepsilon| \le \bar{\varepsilon_0}$), we obtain

$$\begin{aligned} \|u_\lambda\|_p^p &\ge |\varepsilon|^p \tilde{C}\eta/\sqrt{|\lambda-\lambda_0|} \ge |\bar{\varepsilon_0}|^p \tilde{C}\eta/\sqrt{|\lambda-\lambda_0|} \\ &\ge C/\sqrt{|\lambda-\lambda_0|}. \qquad \square \end{aligned}$$

Remark. Suppose, in addition, that (H3) holds. If $T(T_1+T_{-1})<0$ and $T(S_1+S_{-1})<0$ (i.e., $T_1+T_{-1}>0$ and $S_1+S_{-1}>0$ in case $\lambda_0 \in \partial_\ell\sigma$ or $T_1+T_{-1}<0$ and $S_1+S_{-1}<0$ in case $\lambda_0 \in \partial_r\sigma$) then, by Corollary 4.5, $M(\varepsilon) \le 0$ ($|\varepsilon| \le \bar{\varepsilon_0}$) for some $\bar{\varepsilon_0} \in (0,\varepsilon_0)$. With respect to Example 4.6, $M(\varepsilon) \le 0$ ($|\varepsilon| \le \bar{\varepsilon_0}$) is realized if either $\lambda_0 \in \partial_r\sigma$ and $r_1 \le 0$ or $\lambda_0 \in \partial_\ell\sigma$ and $r_1 \ge 0$.

A typical situation for Theorem 5.1 is illustrated in Fig. 5.1.

Finally, we treat the case in which $M(\varepsilon) > 0$ ($\varepsilon_2 < \varepsilon < \varepsilon_3$) for some $\varepsilon_2, \varepsilon_3 \in [-\varepsilon_0, \varepsilon_0]$. We obtain solutions of the original problem and we are able to discuss their bifurcation. In particular, for even nonlinearities, for example, we can get bifurcation from 0 at both ends of a gap by use of Example 4.7—a result which could not be obtained by variational methods so far.

THEOREM 5.2. *Suppose that* (H0), (H1), *and* (H2) *hold. Assume that* $(v_\varepsilon, M(\varepsilon))$ *are solutions of* (4.5), (4.6) *as in Theorem* 4.3, *satisfying* $0 < M(\varepsilon)$ $(\varepsilon \in (\varepsilon_2,\varepsilon_3) \subset [-\varepsilon_0,\varepsilon_0])$. *Then* (u_λ,λ), *given by* (5.1), (5.2), *forms a continuous branch of nontrivial solutions of* (1.1) *such that*

$$\|u_\lambda\|_p^p = \frac{|\varepsilon|^p}{M(\varepsilon)} \frac{1}{p\sqrt{|D'(\lambda_0)|}} \left[\int_{-N}^{N} |\psi(x)|^p dx (1+|1+c(\varepsilon)|^p) + M(\varepsilon)\xi(\varepsilon) \right], \tag{5.4}$$

where $\xi : [\varepsilon_2,\varepsilon_3] \to \mathbb{R}$ *is a continuous function.*

Remarks. (i) Suppose $\varepsilon_2 = 0$ or $\varepsilon_3 = 0$.

(a)Then u_λ is continuous in λ_0 with $u_{\lambda_0} = 0$ if and only if $\|u_\lambda\|_p^p \to 0$ for $\lambda = \lambda_0 + s_0 M^2(\varepsilon)$, $\varepsilon \to 0$.

(b) If there are $C>0$ and $\tau>0$ such that $M(\varepsilon) \ge C|\varepsilon|^\tau$ ($\varepsilon \in [\varepsilon_2,\varepsilon_3]$), then, for $\varepsilon \to 0$,

$$\|u_\lambda\|_p^p \le \frac{4}{Cp\sqrt{|D'(\lambda_0)|}} |\varepsilon|^{p-\tau} \int_{-N}^{N} |\psi(x)|^p dx, \tag{5.5}$$

hence there is bifurcation from 0 at λ_0 if $\tau < p$.

(c) If there are $C>0$ and $\tau>0$ such that $0 < M(\varepsilon) \le C|\varepsilon|^\tau$ ($\varepsilon \in [\varepsilon_2,\varepsilon_3]$), then, for $\varepsilon \to 0$,

$$\|u_\lambda\|_p^p \ge \frac{|\varepsilon|^{p-\tau}}{Cp\sqrt{|D'(\lambda_0)|}} \int_{-N}^{N} |\psi(x)|^p dx, \tag{5.6}$$

hence there is asymptotic bifurcation from infinity if $\tau > p$.

(ii) Suppose $\varepsilon_2 \ne 0 \ne \varepsilon_3$ and $M(\varepsilon_2) = 0$ or $M(\varepsilon_3) = 0$. Then there exists $C>0$ such that, for $\varepsilon \in (\varepsilon_2,\varepsilon_3)$,

$$\|u_\lambda\|_p^p \ge C/M(\varepsilon).$$

(iii) Suppose, in addition, that (H3) holds.

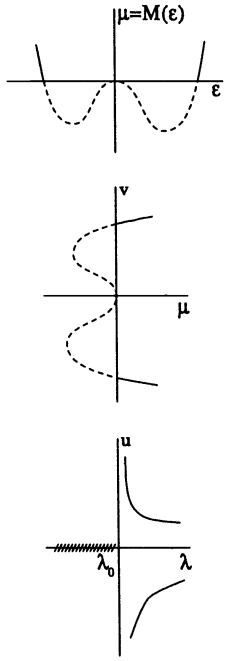


FIG. 5.1. *In the case of a subcritical bifurcation for v there is (locally) no branch for u; if the v branch, however, turns back as above, this leads to a new u branch bifurcating from infinity. (Note that only the solid lines in the (μ, v) diagram lead to solutions in the (λ, u) diagram.)*

(a) If $T(T_1 + T_{-1}) > 0$ (i.e., $T_1 + T_{-1} > 0$ and $\lambda_0 \in \partial_r\sigma$ or $T_1 + T_{-1} < 0$ and $\lambda_0 \in \partial_\ell\sigma$) then, by Corollary 4.5,

$$\|u_\lambda\|_p^p = T_0\varepsilon^{p-\sigma}\left(\int_{-N}^{N} |\psi(x)|^p dx + \xi^+(\varepsilon)\right) \quad (0 < \varepsilon \le \varepsilon_3),$$

where

$$T_0 = \frac{2T}{p\sqrt{|D'(\lambda_0)|}(T_1 + T_{-1})}$$

and $\xi^+ : [0, \varepsilon_3] \to \mathbb{R}$ is continuous and $\xi^+(0) = 0$.

(b) If $T(S_1 + S_{-1}) > 0$ (i.e., $S_1 + S_{-1} > 0$ and $\lambda_0 \in \partial_r\sigma$ or $S_1 + S_{-1} < 0$ and $\lambda_0 \in \partial_\ell\sigma$) then, by Corollary 4.5,

$$\|u_\lambda\|_p^p = S_0|\varepsilon|^{p-\sigma}\left(\int_{-N}^{N} |\psi(x)|^p dx + \xi^-(\varepsilon)\right) \quad (\varepsilon_2 \le \varepsilon < 0),$$

where

$$S_0 = \frac{2T}{p\sqrt{|D'(\lambda_0)|}(S_1 + S_{-1})}$$

and $\xi^- : [\varepsilon_2, 0] \to \mathbb{R}$ is continuous and $\xi^-(0) = 0$.

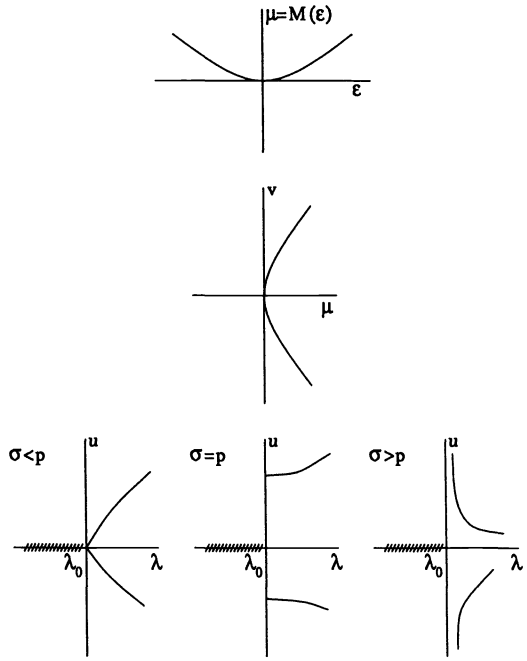


FIG. 5.2. *According to the growth of the nonlinearity, pitchfork bifurcation for v is transformed into a kind of pitchfork "bifurcation" for u either from 0, a finite number, or infinity.*

Hence, in both cases we obtain a bifurcation behavior which has previously been observed for bifurcation at the lowest point of the continuous spectrum [18], [22], [23]:

(α) There is bifurcation from 0 at $\lambda = \lambda_0$ if $\sigma < p$.

(β) There is bifurcation from $T_0 \cdot I$ or $S_0 \cdot I$ at $\lambda = \lambda_0$ if $\sigma = p, I = \int_{-N}^{N} |\psi(x)|^p dx$.

(γ) There is bifurcation from infinity at $\lambda = \lambda_0$ if $\sigma > p$.

Typical forms of bifurcation are illustrated in Figs. 5.1–5.3; special situations can be realized by choosing coefficients and exponents appropriately in Examples 4.6 and 4.7.

Proof. Using (5.1), (5.2), and Lemma 3.2, we calculate

$$\|u_\lambda\|_p^p = \int_{-N}^{N} |v_\varepsilon(x)|^p dx + \frac{1}{1-|\varrho(\lambda)|^p} \left[|c_{-1}(\lambda)|^p |\varrho(\lambda)|^p \int_{-N}^{N} |u_{-1}(x,\lambda)|^p dx \right.$$
$$\left. + |c_1(\lambda)|^p \int_{-N}^{N} |u_1(x,\lambda)|^p dx \right]$$
$$= |\varepsilon|^p \int_{-N}^{N} |\psi + \gamma(\varepsilon)|^p(x) dx + \frac{|\varepsilon|^p}{1-|\varrho(\lambda)|^p} \left[|\varrho(\lambda)|^p \int_{-N}^{N} |u_{-1}(x,\lambda)|^p dx \right.$$
$$\left. + |1 + c(\varepsilon)|^p \int_{-N}^{N} |u_1(x,\lambda)|^p dx \right].$$

By the analyticity of $u_{-1}(x,\lambda)$, $u_1(x,\lambda)$ with respect to λ and using $u_{-1}(x,\lambda_0) = u_1(x,\lambda_0) = \psi(x)$, there exist continuous functions $t_{-1}, t_1 : [0, \mu_0^2] \to \mathbb{R}$ such that, for

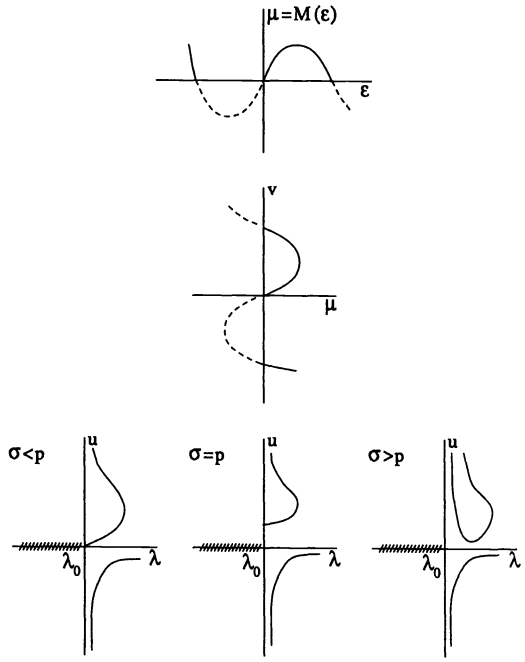


FIG. 5.3. *Transcritical bifurcation for v leads to a single branch for u bifurcating either from 0, a finite number, or infinity.*

all $\lambda \notin \sigma$, $|\lambda - \lambda_0| \leq \mu_0^2$,

$$\int_{-N}^{N} |u_i(x,\lambda)|^p dx = \int_{-N}^{N} |\psi(x)|^p dx + t_i(|\lambda - \lambda_0|)\,|\lambda - \lambda_0| \quad (i = -1, 1).$$

Hence

$$\|u_\lambda\|_p^p = \frac{|\varepsilon|^p}{M(\varepsilon)} \frac{1}{p\sqrt{|D'(\lambda_0)|}} \left[\int_{-N}^{N} |\psi(x)|^p dx (1 + |1 + c(\varepsilon)|^p) + M(\varepsilon)\xi(\varepsilon) \right],$$

where

$$\begin{aligned}\xi(\varepsilon) = {} & p\sqrt{|D'(\lambda_0)|} \int_{-N}^{N} |\psi(x) + \gamma(\varepsilon)(x)|^p dx \\ & + \frac{\vartheta(M(\varepsilon))(1+|1+c(\varepsilon)|^p) - p^2|D'(\lambda_0)| + p\sqrt{|D'(\lambda_0)|}\vartheta(M(\varepsilon))M(\varepsilon)}{p\sqrt{|D'(\lambda_0)|} - \vartheta(M(\varepsilon))M(\varepsilon)} \int_{-N}^{N} |\psi(x)|^p dx \\ & + \frac{p\sqrt{|D'(\lambda_0)|}M(\varepsilon)}{p\sqrt{|D'(\lambda_0)|} - \vartheta(M(\varepsilon))M(\varepsilon)} (t_{-1}(M^2(\varepsilon))|\tilde{\varrho}(M(\varepsilon))|^p + t_1(M^2(\varepsilon))|1 + c(\varepsilon)|^p),\end{aligned}$$

where ϑ is as in the proof of Theorem 5.1.

The continuity of u_λ with respect to λ (resp., ε) is a direct consequence of the continuous dependence of $\lambda, v_\varepsilon, c(\varepsilon)$ of ε.

REFERENCES

[1] S. ALAMA, *An eigenvalue problem and the color of crystals*, Ph.D. thesis, New York University, 1988.

[2] S. ALAMA AND Y. Y. LI, *Existence of solutions for semilinear elliptic equations with indefinite linear part*, J. Differential Equations, 96 (1992), pp. 89–115.

[3] M. S. P. EASTHAM, *The Spectral Theory of Periodic Differential Equations*, Scottish Academic Press, Edinburgh, 1973.

[4] H. P. HEINZ, *Bifurcation from the essential spectrum for nonlinear perturbations of Hill's equation*, in Differential Equations—Stability and Control, S. Elaydi, ed., Marcel Dekker, New York, 1990, pp. 219–226.

[5] ———, *Lacunary bifurcation for operator equations and nonlinear boundary value problems on* $\mathbb{R}^n$, Proc. Roy. Soc. Edinburgh Sect. A, 118 (1991), pp. 237–270.

[6] ———, *Lacunary bifurcation of multiple solutions of nonlinear eigenvalue problems*, in Internat. Ser. Numer. Math. 97, Birkhäuser, Basel, 1991, pp. 161–169.

[7] ———, *Existence and gap-bifurcation of multiple solutions to certain nonlinear eigenvalue problems*, Nonlinear Anal., 21 (1993), pp. 457–484.

[8] ———, *On the number of solutions of nonlinear schrödinger equations and on unique continuation*, J. Differential Equations, 116 (1995), pp. 149–171.

[9] H. P. HEINZ, T. KÜPPER, AND C. A. STUART, *Existence and bifurcation for nonlinear perturbations of the periodic Schrödinger equation*, J. Differential Equations, 100 (1992), pp. 341–354.

[10] H. P. HEINZ AND C. A. STUART, *Solvability of nonlinear equations in spectral gaps of the linearization*, Nonlinear Anal., 19 (1992), pp. 124–164.

[11] T. KÜPPER, *Verzweigung aus dem wesentlichen Spektrum*, GAMM-Mitteilungen, Heft 1 (1991), pp. 11–22.

[12] T. KÜPPER AND C. A. STUART, *Bifurcation into gaps in the essential spectrum*, J. Reine Angew. Math., 409 (1990), pp. 1–34.

[13] ———, *Gap-bifurcation for nonlinear perturbations of Hill's equation*, J. Reine Angew. Math., 410 (1990), pp. 23–52.

[14] ———, *Bifurcation at Boundary Points of the Continuous Spectrum*, Progr. Nonlinear Differential Equations, Nonlinear Diffusion Equations and Their Equilibrium States 3, Birkhäuser, Boston, 1992.

[15] ———, *Necessary and sufficient conditions for gap-bifurcation*, Nonlinear Anal., 18 (1992), pp. 893–903.

[16] T. MRZIGLOD, *Unbounded solution components for nonlinear Hill's equations*, Proc. Roy. Soc. Edinburgh, to appear.

[17] P. RABINOWITZ, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[18] H. J. RUPPEN, *Inherited bifurcation*, Nonlinear Anal., 19 (1992), pp. 993–1000.

[19] C. A. STUART, *Global properties of components of solutions of non-linear second order ordinary differential equations on the half-line*, Ann. Scuola Norm. Sup. Pisa, 2 (1975), pp. 256–286.

[20] ———, *Some bifurcation theory for k-set contractions*, Proc. London Math. Soc. (3), 27 (1973), pp. 531–550.

[21] ———, *Bifurcation for variational problems when the linearisation has no eigenvalues*, J. Funct. Anal., 38 (1980), pp. 169–187.

[22] ———, *Bifurcation for Neumann problems without eigenvalue*, J. Differential Equations, 36 (1980), pp. 391–407.

[23] ———, *Bifurcation for Dirichlet problems without eigenvalues*, Proc. London Math. Soc. Ser. 3, 45 (1982), pp. 169-192.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1306–1330, September 1995

012

STABILITY CRITERIA FOR SECOND-ORDER DYNAMICAL SYSTEMS INVOLVING SEVERAL TIME DELAYS*

F. G. BOESE†

Abstract. Characteristic functions $F(z)$ for second-order difference-differential equations with constant parameters of the form

$$F(z) := z^2 + zA(z) + B(z), \qquad z \in \mathbf{C},$$
$$A(z) := \sum_{k=0}^{n} A_k e^{-zT_k}, \qquad A_k \in \mathbf{R}, \qquad T_0 := 0, \quad T_k \in \mathbf{R}_+, \quad k = 1(1)n,$$
$$B(z) := \sum_{k=0}^{n} B_k e^{-zT_k}, \qquad B_k \in \mathbf{R}, \qquad n \in \mathbf{N}$$

are studied. When studying delay-independent stability, the class of $F(z)$ is enlarged;

$$F_T(z) := z^2 + A_1 z + A_0 + B(z), \qquad A_1, A_0 > 0, \qquad T := (T_1, \ldots, T_n) \geq 0,$$
$$B(z) := \sum_{k=1}^{n} B_k \cdot b_{k,p_k}(z) e^{-zT_k}, \qquad B_k \in \mathbf{R},$$

where the $b_{k,p_k}(z)$ real monic polynomials of degree $p_k \leq 2$ are considered. For such $F_T(z)$, an explicit subset $\mathcal{S}_0$ of the set of all $F_T(z)$ that are stable for all delay vectors $T \geq 0$ is derived.

Key words. delay systems, time delays, characteristic functions, asymptotic stability

AMS subject classifications. Primary, 34K20; Secondary, 92A17

1. Problem and motivation.

There are several ways to investigate the local asymptotical stability of equilibria of dynamical systems. When only stability against small and time-limited perturbations of the system are considered, the way via characteristic functions is the method of choice. Within this method, one looks for elementary solutions of the form $x_z(t) := e^{zt}$, $t \geq 0$. A function $F(z)$, $z \in \mathbf{C}$, whose zero set Σ, also called *spectrum*, determines the solution set $\{x_z(t)\}_{z \in \Sigma}$, is called a *characteristic function*. To be more specific, we consider the dynamical systems

$$\dot{x}(t) = f\left[x(t-\tau_0), x(t-\tau_1), \ldots, x(t-\tau_m)\right], \qquad t \in \mathbf{R}_+, \qquad m \in \mathbf{N} \tag{1.1}$$

with vector state $x \in \mathbf{R}^d$ and constant delays $\tau_0 := 0$, $\tau_k \in \mathbf{R}_+$, $k = 1(1)m$. Here and in what follows, the overdot stands for differentiation with respect to t. The function $f \in C^1(\mathbf{R}^{md+d}, \mathbf{R}^d)$ is assumed to have *equilibria* $x_* \in \mathbf{R}^d$, i.e., zeros on the diagonal $f(x_*, \ldots, x_*) = 0$. If x_* is the equilibrium under consideration then the *variational system* for (1.1) is

$$\dot{y} = J_0 y(t-\tau_0) + J_1 y(t-\tau_1) + \cdots + J_m y(t-\tau_m), \tag{1.2}$$

*Received by the editors August 23, 1991; accepted for publication (in revised form) January 31, 1994.

†Ganghoferstraße 81, D-81373 München, Germany (gub@mpe-garching.mpg.de).

where J_k is the Jacobian of f with respect to the kth vector argument of f evaluated at x_*. A perturbation $y(t) := Ye^{zt}$, $Y \in \mathbf{R}^d$, $t \geq 0$, solves (1.2) if and only if z is a zero of

$$F(z) := \det\left[zI - J_0 e^{-z\tau_0} - J_1 e^{-z\tau_1} - \cdots - J_m e^{-z\tau_m} \right]. \tag{1.3}$$

Here, I is the identity matrix. The function $F(z)$ has the form

$$\begin{aligned} F(z) &:= z^d + z^{d-1}A_1(z) + \cdots + A_d(z), \\ A_k(z) &:= \sum_{j=0}^{n} A_{k,j} e^{-zT_j}, \qquad A_{k,j} \in \mathbf{R}, \qquad T_k \in \mathbf{R}_+ \cup \{0\}, \end{aligned} \tag{1.4}$$

where the T_k are linear combinations of the τ_k. In dimension $d = 2$, to which we restrict ourselves exclusively from now on, we prefer the simpler notation

$$\begin{aligned} F(z) &:= z^2 + zA(z) + B(z), \\ A(z) &:= \sum_{k=0}^{n} A_k e^{-zT_k}, \qquad A_k \in \mathbf{R}, \qquad T_0 := 0, \qquad T_k \in \mathbf{R}_+, \\ B(z) &:= \sum_{k=0}^{n} B_k e^{-zT_k}, \qquad B_k \in \mathbf{R}, \qquad n \in \mathbf{N}. \end{aligned} \tag{1.5}$$

We name yet another instance which arises to consider the class of exponential polynomials $F(z)$ defined by (1.5).

Consider the one-dimensional forced harmonic oscillator with (positive or negative) damping

$$\ddot{x}(t) + A\dot{x}(t) + Bx(t) = u(t), \quad x(0) := x_0, \quad \dot{x}(0) := \dot{x}_0, \quad A, B, x_0, \dot{x}_0 \in \mathbf{R}, \quad t \in \mathbf{R}_+. \tag{1.6}$$

Evidently, $x_* := 0$ is the sole rest position of the free oscillator $u :\equiv 0$. So, $x(t)$ is, directly, the deviation from the equilibrium. We wish to select a control $u(t)$ from the class (1.7) such that the initial perturbation x_0, $\dot{x}_0$ dies out fast and with no oscillation. To achieve this goal, linear proportional-derivative state feedback may be implemented;

$$u(t) := -\sum_{k=0}^{n} A_k \dot{x}(t - T_k) - \sum_{k=0}^{n} B_k x(t - T_k). \tag{1.7}$$

The characteristic function $F(z)$ for the closed loop system which results when substituting (1.7) in (1.6) is $F(z)$ from (1.5). Also, if retarded derivatives of second order are allowed in $u(t)$, then we have to add the term

$$u_2(t) := -\sum_{k=0}^{n} C_k \ddot{x}(t - T_k) \tag{1.8}$$

to the right-hand side of (1.7), which gives, with $A(z)$, $B(z)$ from (1.5),

$$\begin{aligned} F(z) &:= z^2 C(z) + zA(z) + B(z), \\ C(z) &:= 1 + \sum_{k=0}^{n} C_k e^{-zT_k}, \qquad C_k \in \mathbf{R}. \end{aligned} \tag{1.9}$$

The last $F(z)$ is *neutral* for $|C_1|+|C_2|+\cdots+|C_n|>0$. The finite-dimensional version of (1.6) is written as follows:

$$\begin{aligned} &\dot{x}(t)+Bx(t)=Cu(t), \qquad B\in\mathbf{R}^{d\times d}, \qquad C\in\mathbf{R}^{d\times m},\\ &u(t):=-\sum_{k=0}^{n}\{A_k\dot{x}(t-T_k)+B_k(t-T_k)\}, \qquad A_k,B_k\in\mathbf{R}^{m\times d}. \end{aligned} \tag{1.10}$$

The characteristic function $F(z)$ belonging to (1.10) is

$$\begin{aligned} F(z)&:=\det[\,zA(z)+B(z)\,],\\ A(z)&:=I+C\cdot\left\{\sum_{k=0}^{n}A_ke^{-zT_k}\right\},\\ B(z)&:=B+C\cdot\left\{\sum_{k=0}^{n}B_ke^{-zT_k}\right\}. \end{aligned} \tag{1.11}$$

We observe that $F(z)$ from (1.11) is, for $d=2$, again of the form (1.5) or (1.9). Consider (1.10) for $A_k:=0$, $k=1(1)n$, and with vanishing delays. Then the pole shifting theorem, the backbone of linear control theory, applies under the generic condition of controllability and says that the control goal from above can be reached. For positive delays, however, such a theorem is not at our disposal. So, the treatment of $F(z)$ from (1.5) also has a bearing on linear control systems in dimesion $d=2$.

We may parameterize the class (1.5) by introducing the real $(3n+2)$ vector

$$p:=(A_0,\ldots,A_n,B_0,\ldots,B_n,T_1,\ldots,T_n) \tag{1.12}$$

and consider p in the *coefficient parameter space*

$$p\in\mathbf{P}:=\mathbf{P}_0\times\mathbf{T}, \qquad \mathbf{P}_0:=\mathbf{R}^{2n+2}, \qquad \mathbf{T}:=\mathbf{R}_+^n, \tag{1.13}$$

where $\mathbf{P}_0$ is spanned by the coefficients A_0 to B_n and $\mathbf{T}$ is spanned by the delays T_1 to T_n. We are faced with two problems: the determination of the *stability chart*

$$\mathcal{S}:=\{\,p\in\mathbf{P}:\ F(z)\neq 0 \text{ in } \operatorname{Re}(z)\geq 0\,\} \tag{1.14}$$

and the development of *stability criteria*. While a stability chart lists all stable $F(z)$, a stability criterion decides whether or not a single $F(z)$ is stable. Extreme robustness against variation of the delays is embodied in the notion of *stability independent of delays*. We split the full parameter vector p into two direct summands $p_0:=(A_0,\ldots,B_n)$ and $T:=(T_1,\ldots,T_n)$ so that $p:=(p_0,T)$ with $p_0\in\mathbf{P}_0$ and $T\in\mathbf{T}$. The *stability chart $\mathcal{S}_0$, independent of the delays T* is defined by

$$\mathcal{S}_0:=\{\,p_0\in\mathbf{P}_0:\ F(z)\neq 0 \text{ in } \operatorname{Re}(z)\geq 0 \text{ for all } T\in\mathbf{T}\,\}. \tag{1.15}$$

In §2, we derive a stability criterion which covers the whole class (1.5). Section 3 contains an example. Section 4 is devoted to the stability independent of delays. The concluding section, §5, contains a discussion.

The letter i is exclusively reserved for the imaginary unit, $i^2=-1$. The expression $k=a(b)c$ means variable k runs from a to c in steps of b. $A:=B$ or $B=:A$ redefines A by B. $\operatorname{Arg}(z)$ denotes the general argument function and $\arg(z)$ is its principal

branch with values in $(-\pi, \pi]$, where $\arg(z) = \pi$ on the negative real axis, which is the branch cut.

2. A stability criterion for a special class of holomorphic functions. As already mentioned in §1, it is widely known that necessary and sufficient stability criteria for holomorphic functions $F(z)$ for one complex variable z are, in general, not simple.

A more general version, in several respects, of our criterion can be established; see [3] or the recent book of Stépán [16]. We disregard greatest generality and tailor the theorem for the application we envisage.

In preparation of the statement and proof of the theorem, we introduce two definitions and some notation used in what follows.

Let $s := \{ s_1, s_2, \dots, s_m \}$, $m \in \mathbf{N}_0$, be a finite sequence of nonvanishing numbers $s_k \in \mathbf{R} \setminus \{0\}$. Then we define the set $\mathcal{S}(s)$ of sign changing indices in s or, in short, the *sign changes* of s, by

$$\mathcal{S}(s) := \{ k \in \{ 1, 2, \dots, m-1 \} : \ s_k s_{k+1} < 0 \}. \tag{2.1}$$

By definition, $m-1$ is the largest possible sign change in s. There is a continuous counterpart. Let $f \in C(\mathbf{R}, \mathbf{R})$ be continuous. Then $y \in \mathbf{R}$ is a *sign change* of f if $f(y-\epsilon)f(y+\epsilon) < 0$ for all $\epsilon > 0$ sufficiently small. We turn to some notation. The imaginary axis in $\mathbf{C}$, viewed as an oriented curve, is denoted by I when upwards oriented and $-I$ when downwards oriented. The parts of I falling in $\mathrm{Im}(z) \geq 0$ or $\mathrm{Im}(z) \leq 0$ are I_+ or I_-, respectively. The part of I belonging to the horizontal strip $-R \leq \mathrm{Im}(z) \leq R$, $R > 0$, is designated by I_R. By $[a, b]$, we denote the oriented line segment from a to b, $a, b \in \mathbf{C}$. The semicircle of $|z| = R$ lying in $\mathrm{Re}(z) \geq 0$ with initial point $z = -iR$ and terminal point $z = iR$ is denoted by S_R. Finally, consider the oriented Jordan arc C_{z_1,z_2} with initial point $z = z_1$ and terminal point $z = z_2$ such that $0 \notin C_{z_1,z_2}$. Then

$$\Delta(C_{z_1,z_2}) := \mathrm{Arg}(z_2) - \arg(z_1) \tag{2.2}$$

denotes the *argument variation*, i.e., the variation of $\arg(z)$ when z traces C_{z_1,z_2} in its orientation. The branch of $\mathrm{Arg}(z)$ is chosen so that it varies continuously with z along the curve. We shall see that the evaluation of Δ, i.e., the choice of the correct branch of $\mathrm{Arg}(z_2)$, is the most difficult step.

THEOREM 2.1. *Let $F(z)$ be a real, holomorphic function defined in an open set $\mathcal{O}$ containing the closed right half-plane $\mathrm{Re}(z) \geq 0$ having the following properties*

$$\begin{aligned} &(1) \quad F(0) > 0, \\ &(2) \quad F(iy) \neq 0, \quad y \in \mathbf{R}, \\ &(3) \quad F(z) = z^{2n}[1 + o(1)], \quad |z| \to \infty, \quad \mathrm{Re}(z) \geq 0, \quad n \in \mathbf{N}. \end{aligned} \tag{2.3}$$

Let

$$u_1 < u_2 < \cdots < u_m, \qquad m \in 1 + 2\mathbf{N} \tag{2.4}$$

be the sequence of sign changes of $U(y) := \mathrm{Re}[F(iy)]$, $y > 0$. Denote by σ the sign sequence

$$\begin{aligned} \sigma &:= \{ \sigma_1, \sigma_2, \dots, \sigma_m \}, \\ \sigma_k &:= \mathrm{sgn}\{ \mathrm{Im}[F(iu_k)] \}. \end{aligned} \tag{2.5}$$

Then $F(z) \neq 0$ *in* $\mathrm{Re}(z) \geq 0$ *if and only if*

$$n = \frac{\sigma_1 + \sigma_m}{2} - \sum_{k \in \mathcal{S}} (-1)^k \sigma_k, \tag{2.6}$$

where $\mathcal{S} := \mathcal{S}(\sigma)$ *is the sequence of sign changes in* σ.

Proof. Consider the closed, simple, positively oriented family of curves

$$C_R := -I_R \cup S_R \tag{2.7}$$

for all $R > 0$. Given a point z_0 in $\mathrm{Re}(z) > 0$, one can always choose R such that z_0 is inside C_R. Besides the preimage z plane, we introduce a w plane as image plane and consider $C'_R := F(C_R)$ there. Here, and in what follows, the prime stands for objects in the image plane. The *argument principle* of the theory of functions tells us that the number N of zeros of $F(z)$ in C_R equals the *winding number* W of C'_R with respect to the origin $w = 0$ of the image plane. The winding number (see Ahlfors [1, p. 114]) is proportional to the the argument variation $W := \Delta(C'_R)/(2\pi)$, and we may express N in terms of Δ as

$$N := \frac{\Delta(C'_R)}{2\pi}. \tag{2.8}$$

The obvious additivity of the argument variation Δ with respect to the curve arc and the fact that $\Delta(-I') = -\Delta(I')$ allow us to write (2.8) after the passage to the limit $R \to \infty$ as

$$\begin{aligned} N &= \frac{-\Delta(I'_-) - \Delta(I'_+) + \lim\{\Delta(S'_R) : \ R \to \infty\}}{2\pi}, \\ &= n - \frac{\Delta(I'_+)}{\pi}. \end{aligned} \tag{2.9}$$

Condition (3) of (2.3) tells us that N and all quantities on the right-hand side of (2.9) have limits for $R \to \infty$. The representation of $F(z)$ in (3) of (2.3) directly allows us to read off the last limit in (2.9) as $2n\pi$. From the reality of $F(z)$, $\Delta(I'_-) = \Delta(I'_+)$ follows so that the last line of (2.9) results. So, the only nontrivial part of the proof is the finding of a representation for $\Delta(I'_+)/\pi$ in finitely many terms. To this end, we decompose $F(iy)$, $y \geq 0$, in real and imaginary parts

$$F(iy) =: U(y) + iV(y). \tag{2.10}$$

By (1) and (3) from (2.3), we are informed that $U(0) > 0$ and $(-1)^n U(y) \to +\infty$ for $y \to +\infty$. Hence, $U(y)$ has an odd number of sign changes $m \in 1 + 2\mathbf{N}_0$ for odd n and an even number $m \in 2\mathbf{N}_0$ for even n,

$$u_0 := 0 < u_1 < u_2 < \cdots < u_m < +\infty =: u_{m+1}. \tag{2.11}$$

The outer members on (2.11) are not sign changes but this notation is useful as we shall see in a moment. We take the sign changes of (2.12) as partition points for I_+ and decompose

$$\begin{aligned} I_+ &:= I_0 \cup I_1 \cup \cdots \cup I_m, \\ I_k &:= [iu_k, iu_{k+1}], \qquad k = 0(1)m. \end{aligned} \tag{2.12}$$

The intervals I_k inherit their orientation from I_+. Inserting (2.12) in the last line of (2.9) gives

$$N = n - \frac{\Delta(I_0') + \Delta(I_m')}{\pi} - \sum_{k=1}^{m-1} \frac{\Delta(I_k')}{\pi}. \tag{2.13}$$

The determination of the Δ in (2.13) now poses no problem. We first observe that

$$\operatorname{sgn}[U(y)] = (-1)^k, \qquad u_k < y < u_{k+1}, \qquad U(y) \neq 0. \tag{2.14}$$

Note well that $U(y)$ may have zeros in the interior of $[u_k, u_{k+1}]$ but no sign changes. Condition (2) of (2.3) guarantees that no arc I_k' contains the origin $w = 0$. According to (2.14) it lies entirely in $\operatorname{Re}(z) \geq 0$ for even k and $\operatorname{Re}(z) \leq 0$ for odd k. The initial point of I_k is $w = iV_k$, $V_k := V(u_k)$, and $w = iV_{k+1}$ is the terminal point. Since $V_k \neq 0$ for $k = 1(1)m$, no sign $\sigma_k := \operatorname{sgn}[V_k]$ vanishes. In terms of the signs σ_k, by an elementary case distinction with respect to the parity of k and the signs σ_k, σ_{k+1}, one finds

$$\frac{\Delta(I_k')}{\pi} := (-1)^k \cdot \frac{\sigma_{k+1} - \sigma_k}{2}, \qquad k = 1(1)m - 1. \tag{2.15}$$

The argument variation along the first and last image arc is found as

$$\frac{\Delta(I_0')}{\pi} := \frac{\sigma_1}{2}, \qquad \frac{\Delta(I_m')}{\pi} := \frac{\sigma_m}{2}. \tag{2.16}$$

Inserting (2.15), (2.16) in (2.13) results in

$$N = n - \frac{\sigma_1 + \sigma_m}{2} + \sum_{k=1}^{m-1} (-1)^k \cdot \frac{\sigma_k - \sigma_{k+1}}{2}. \tag{2.17}$$

Only the sign-changing indices k in $\sigma := \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ contribute to the sum in (2.17). Hence,

$$N = n - \frac{\sigma_1 + \sigma_m}{2} + \sum_{k \in \mathcal{S}} (-1)^k \sigma_k. \tag{2.18}$$

Finally, solving $N = 0$ for n gives the necessary and sufficient condition

$$n = \frac{\sigma_1 + \sigma_m}{2} - \sum_{k \in \mathcal{S}} (-1)^k \sigma_k, \tag{2.19}$$

which is the assertion of the theorem. □

At the expense of introducing new quantities, (2.6) may be written in a simpler form. Let $S := |\mathcal{S}|$ be the number of sign changes in σ, S_o be the number of sign changes on odd indices, and S_e, those on even indices. Then clearly,

$$S = S_o + S_e. \tag{2.20}$$

In terms of these quantities, (2.6) gains the shape

$$\begin{aligned} n &= \frac{\sigma_1 + \sigma_m}{2} - S_o + S_e, \\ &= \frac{\sigma_1 + \sigma_m}{2} + S - 2S_o. \end{aligned} \tag{2.21}$$

Given n, there must be enough sign changes in σ to satisfy (2.6). The right-hand side of (2.6) attains its maximum m at the alternating sequence $\sigma_a := \{1, -1, 1, \ldots, 1\}$ with an odd number m of elements. Therefore

$$m \geq n, \qquad S \geq n - 1 \tag{2.22}$$

are necessary conditions for 2.6 to hold.

Consider the real polynomials

$$F(z) := z^{2n} + a_{2n-1} z^{2n-1} + \cdots + a_0. \tag{2.23}$$

Here, (2.6) and (2.22) lead to the stronger condition $\sigma = \sigma_a$, which means, geometrically, that the zeros of $U(y)$ and $V(y)$ are interlaced and $U(y)$ and $V(y)$ have no nonreal zeros. The Hermite–Biehler theorem (see Obreschkoff [13, p. 13] or Lewin [10, p. 305]) tells us that polynomials $F(z)$ of odd degree also share this property. It can even be formulated in classes of entire transcendental functions; cf. Lewin [10, pp. 311–327]. We shall come back to that point in §5.

The $F(z)$ in (5.1) belong to the case $n = 1$ in Theorem 2.1. We may rewrite (2.21) as

$$\begin{aligned} 1 &= \sigma_1 + \omega, \\ \omega &:= \frac{\sigma_m - \sigma_1}{2} + S_e - S_o. \end{aligned} \tag{2.24}$$

Denote by $\mathcal{F}$ the exact class of functions $F(z)$ from (1.5) for which $\omega = 0$ holds true. Are there $F(z)$ which do not belong to $\mathcal{F}$? Trivially, the case $m = 1$, i.e., functions $F(z)$ for which $U(y)$ possesses only one zero, belongs to $\mathcal{F}$. In the next, less simple case, $U(y)$ has exactly $m = 3$ zeros. Now (2.6), viewed as a Diophantine equation for $\sigma := \{\sigma_1, \sigma_2, \sigma_3\}$, may be rewritten as

$$1 = \frac{\sigma_1 + \sigma_3}{2} + \begin{cases} 0, & \mathcal{S} = \emptyset, \\ \sigma_1, & \mathcal{S} = \{1\}, \\ -\sigma_2, & \mathcal{S} = \{2\}. \end{cases} \tag{2.25}$$

So, the sign sequences solving (2.6) for $n = 1$ and $m = 3$ are

$$\sigma \in \{(1,1,1), (1,-1,-1), (-1,-1,1)\} =: \Sigma_3. \tag{2.26}$$

These are the sequences which have up to one sign change. The sign sequence $\sigma := \{1, 1, -1\}$ does not belong to Σ_3. But it is not clear that an $F(z)$ can be found yielding that σ. Next we shall consider a nontrivial example for $m = 7$ for which $\omega \neq 0$.

3. An example. The example serves two purposes: (a) it illustrates the application of Theorem 2.1 and (b) it disproves a claim of Freedman and Rao [5, eq. (3.4)]. These authors consider $F(z)$ of the form

$$\begin{aligned} F(z) &:= z^2 + zA(z) + B(z), \\ A(z) &:= \sum_{k=0}^{3} A_k e^{-zT_k}, \quad A_k \in \mathbf{R}, \quad T_0 := 0, \quad T_1, T_2, T_3 \in \mathbf{R}_+, \\ B(z) &:= \sum_{k=0}^{3} B_k e^{-zT_k}, \quad B_k \in \mathbf{R}, \end{aligned} \tag{3.1}$$

under the restrictions

$$T_3 := T_1 + T_2, \qquad A(0) > 0, \qquad B(0) > 0, \qquad A_3 := 0. \tag{3.2}$$

We decompose $F(iy)$ and the coefficients into real and imaginary parts and write $F(iy) =: U(y) + iV(y)$, $A(iy) =: A'(y) - iA''(y)$, $B(iy) =: B'(y) - iB''(y)$ and find

$$\begin{aligned} U(y) &:= -y^2 + yA''(y) + B'(y), \\ V(y) &:= yA'(y) - B''(y). \end{aligned} \tag{3.3}$$

Any zero $y > 0$ of $U(y)$ fulfills

$$\begin{aligned} y^2 &= yA''(y) + B'(y) \\ &\le yA'' + B', \\ Y &:= \begin{cases} \frac{\sqrt{D}+A''}{2}, & D \ge 0, \\ 0, & D < 0, \end{cases} \\ D &:= A''^2 + 4B', \end{aligned} \tag{3.4}$$

when $A'' \ge A''(y)$, $B' \ge B'(y)$ for all $y \ge 0$. Such constants are

$$\begin{aligned} A'' &:= |A_0| + |A_1| + |A_2| + |A_3|, \\ B' &:= B_0 + |B_1| + |B_2| + |B_3|. \end{aligned} \tag{3.5}$$

We consider the class of functions $F_\epsilon(z)$, $\epsilon := (\epsilon_1, \epsilon_2) \in \mathbf{R}_+^2$, defined by

$$\begin{array}{llll} A_0 := 3 + \epsilon_1, & A_1 := -6, & A_2 := 3, & A_3 := 0, \\ B_0 := 17/2 + \epsilon_2, & B_1 := 3, & B_2 := 3/2 - 2\epsilon_2, & B_3 := 3, \\ T_0 := 0, & T_1 := \pi/2, & T_2 := \pi, & T_3 := 3\pi/2 \end{array} \tag{3.6}$$

for $\epsilon_1 > 0$, $0 < \epsilon_2 < 16$, so that $A(0) > 0$ and $B(0) > 0$, and F_ϵ belongs to the class defined by (3.1), (3.2). In the next lemma, we choose

$$\epsilon_1 := 10^{-6}, \qquad \epsilon_2 := 10^{-3}. \tag{3.7}$$

Under this choice, zeros of $U(y)$ are only possible for $0 < y < Y$. As Table 3.1 reveals, the bound Y overestimates the largest zero by about a factor of 2.

LEMMA 3.1. *The function $F_\epsilon(z)$ defined by* (3.1), (3.6), *and* (3.7) *belongs to the class* (3.2) *and has the following properties*:

(3.8) (*a*) $\operatorname{Im}[F_\epsilon(iy_1)] > 0$, *where y_1 is the smallest positive zero of* $\operatorname{Re}[F(iy)]$;
(*b*) *the number N of zeros of F_ϵ in* $\operatorname{Re}(z) \ge 0$ *is* $N := 2$.

Proof. Evidently, F_ϵ belongs under the choice (3.7) to the subclass (3.2). Next, we have to determine the zeros of $U(y) := \operatorname{Re}[F(iy)]$. This is done numerically. Table 3.1 exhibits the (rounded) values of all seven positive zeros $y = y_k$, $k = 1(1)7$,

TABLE 3.1

Table of the positive zeros y_k, $k = 1(1)7$, *of* $U(y) := \mathrm{Re}[F_\epsilon(iy)]$ *along with the values* $V(y_k) := \mathrm{Im}[F_\epsilon(iy_k)]$ *and the signs* $\sigma_k := \mathrm{sgn}[V(y_k)]$ *for* F_ϵ *from Fig.* 3.1.

k	y	V	σ
1	1.000375	0.005300	1
2	2.000030	24.000778	1
3	3.579722	2.478063	1
4	3.726218	3.418344	1
5	3.999875	0.002950	1
6	6.310892	66.119148	1
7	7.000036	-0.002197	-1

of $U(y)$ along with the values $V(y_k)$ and the sign sequence σ with elements $\sigma_k := \mathrm{sgn}[V(y_k)]$, $k = 1(1)7$.

The reader, equipped with a pocket calculator, may easily verify the correctness of the table entries. It is more cumbersome to show that the table comprises *all* zeros lying in the interval $0 < y < 13.2$. Figures 3.1 and 3.2 clarify this point. The curve disappearing on the left margin of Fig. 3.1 remains in $\mathrm{Re}(W) < 0$. Knowing this, we count that $F(iy)$ intersects $F(W) = 0$ seven times. Figure 3.2 resolves the tiny loop in Fig. 3.1 near the imaginary axis and let us observe that it has two intersection points. Given the correctness of Table 3.1, property (a) of (3.8) is obvious. To see (b) of (3.8), we consider the sequence $\sigma := \{\sigma_1, \ldots, \sigma_7\}$ with elements from the last column of Table 3.1. The set $\mathcal{S}$ of sign changes in σ is $\mathcal{S} := \{6\}$. Application of formula (2.19) yields

$$N = 1 - \frac{\sigma_1 + \sigma_7}{2} + \sum_{k \in \mathcal{S}} (-1)^k \sigma_k = 2, \tag{3.9}$$

which proves the lemma. □

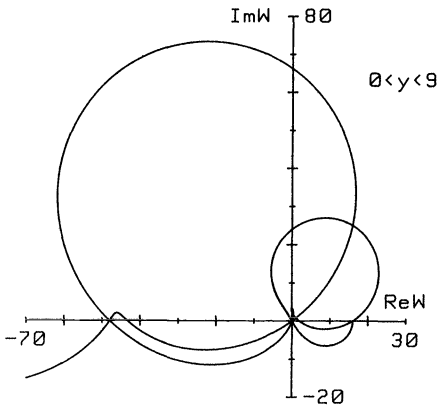


FIG. 3.1. *The arc* $F_\epsilon(iy)$, $0 \le y \le 9$, *in the square* $-70 \le \mathrm{Re}(W) \le 30$, $-20 \le \mathrm{Im}(W) \le 80$ *in the* W *plane for the counterexample* $F_\epsilon(z) := z^2 + z[\epsilon_1 + 3\{1 - e^{-z\pi/2}\}^2] + 17/2 + 6\,[1/4 + \cosh(z\pi/2)]\, e^{-z\pi} + \epsilon_2[1 - 2e^{-z\pi}]$ *with* $\epsilon_1 := 10^{-6}$, $\epsilon_2 := 10^{-3}$.

Property (b) of (3.8) means that $F_\epsilon(z)$ is unstable, so (a) in (3.8) cannot be a correct stability condition in the class considered by Freedman and Rao.

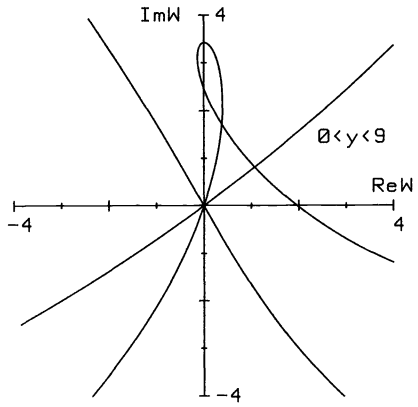


FIG. 3.2. *The part of the arc* $F_\epsilon(iy)$, $0 \le y \le 9$, *falling into the square* $|\mathrm{Re}(W)|, |\mathrm{Im}(z)| \le 4$ *of the* W *plane for* $F_\epsilon(z)$ *from Fig.* 3.1.

Lemma 3.1 does not tell the full truth. The claim of Freedman and Rao is, in a sense, genuinely false. We shall make this clear by showing that the previous counterexample does not hinge on the special choice (3.7).

The function $F_0(z)$ is designed so that $z := iy_k$, $k = 1, 5, 7$, with $y_1 := 1$, $y_5 := 4$, $y_7 := 7$ are zeros of this function. We consider the parameter dependence

$$y_k := y_k(\epsilon_1, \epsilon_2), \qquad k = 1, 5, 7. \tag{3.10}$$

Clearly, $A(z)$, $B(z)$ are analytical functions of ϵ_1, ϵ_2. Hence, $y_k(\epsilon_1, \epsilon_2)$ are piecewise analytical functions and we may expand around the origin $(\epsilon_1, \epsilon_2) = 0$;

$$y_k(\epsilon_1, \epsilon_2) = y_k + y_{k,1}(0,0)\epsilon_1 + y_{k,2}(0,0)\epsilon_2 + o(\|\epsilon\|), \qquad \epsilon \to 0. \tag{3.11}$$

The second indices j, $j = 1, 2$, on the right-hand side of (3.11) stand for partial derivatives with respect to the jth variable of $y_k(\epsilon_1, \epsilon_2)$; $\|\cdot\|$ is a norm in $\mathbf{R}^2_+$. Let the overdot stand for derivation with respect to y; then

$$\dot{U}(y)Y_j(\epsilon_1, \epsilon_2) + U_j(y) = 0, \qquad y_j(0,0) = -\frac{U_j}{\dot{U}(y)}, \qquad j = 1, 2 \tag{3.12}$$

follows from $U(y) = 0$. The right-hand side of the last line is to be evaluated at $\epsilon_1 = \epsilon_2 = 0$. From the definition of $A''(y)$, $y_1(0,0) = 0$ readily follows. We may write

$$B'(y) := B_0'(y) + \epsilon_2 \left[1 - 2\cos(y\pi)\right], \tag{3.13}$$

where $B_0'(y)$ is independent of ϵ_2. So,

$$y_{k,2}(0,0) = \frac{2\cos(y_k\pi) - 1}{\dot{U}(y_k)}. \tag{3.14}$$

Evaluating the last right-hand side after some elementary calculations with positive constants C_k gives

$$y_{k,2}(0,0) = -C_k\epsilon_2 + o(\|\epsilon\|), \qquad C_k > 0, \qquad k = 1, 5, 7, \qquad \epsilon \to 0. \tag{3.15}$$

Next, we expand $V(y)$ at $y = y_k$ and, with $m_1 := 1$, $m_5 := m_7 := 0$, and other positive constants M_k, find

$$\begin{aligned} V(y_k + u) &= V(y_k)u + O(u^2), \\ &= (-1)^{k+m_k} M_k + O(u^2), \qquad M_k > 0, \qquad u \to 0. \end{aligned} \tag{3.16}$$

Inserting (3.14) in (3.15) gives the final asymptotical result for $\epsilon_1 \to 0$, $\epsilon_2 \to 0$:

$$\sigma_k := \operatorname{sgn}[V\{y_k(\epsilon_1, \epsilon_2)\}] = (-1)^{k+m_k}, \qquad k = 1, 5, 7. \tag{3.17}$$

In the scaling of Fig. 3.1, the zeros y_3 and y_4 cannot safely be identified. Therefore, the parts of $F(iy)$, $0 \le y \le 9$, falling in the square $|\operatorname{Re}(W)|, |\operatorname{Im}(W)| \le 4$ are shown in Fig. 3.2. The ordinates of the third and fourth intersection points are in accordance with Table 3.1. Also in this magnification, one cannot read off the ordinates of the first, fifth, and seventh intersection points. A further magnification is necessary. Fig. 3.3 shows the parts of $F(iy)$ in the square $|\operatorname{Re}(W)|, |\operatorname{Im}(W)| \le 0.01$. Now, one clearly sees that the first and fourth ordinates are positive and the seventh is negative. The plot also confirms the first four decimal places behind the decimal point of the ordinate values of Table 3.1 for the intersection points under consideration. When reading Figs. 3.1–3.3, keep in mind that the interior of the images of $\operatorname{Re}(z) \ge 0$ under $F_\epsilon(z)$ lies on the right side of $F(iy)$ when tracing $F(iy)$ in its orientation. This is a trivial consequence of the conformality of the mapping F_ϵ and, properly, not worth mentioning. In our context, however, it is indispensable for a thorough understanding of the plots of $F(iy)$ in terms of coverings of the origin of the image plane. From the foregoing asymptotical analysis, it becomes clear that the pair of zeros of $F_\epsilon(z)$ in $\operatorname{Re}(z) \ge 0$ has the form $z = \pm 7i + \delta$, where δ is a complex number with small modulus and $\operatorname{Re}(\delta) > 0$.

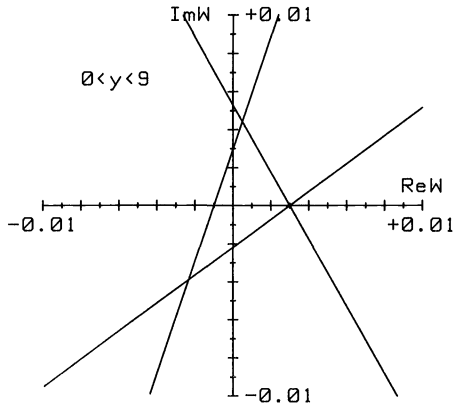


FIG. 3.3. *The part of the arc* $F_\epsilon(iy)$, $0 \le y \le 9$, *falling into the square* $|\operatorname{Re}(W)|, |\operatorname{Im}(z)| \le 0.01$ *of the* W*-plane for* $F_\epsilon(z)$ *from Fig.* 3.1.

4. Delay-independent stability. The largest real parameter space for $F(z)$ from (1.5) is $\mathbf{P} := \mathbf{P}_0 \times \mathbf{T}$ with $\mathbf{P}_0 := \mathbf{R}^{2n+2}$ and $\mathbf{T} := \mathbf{R}^n_+$. $\mathbf{P}_0$, the coefficients space, is spanned by the coefficient vector p of $A(z)$, $B(z)$, and $\mathbf{T}$ is spanned by the delay vector $T := (T_1, \ldots, T_n) \ge 0$. The set $\mathcal{S}_0$ of stability, independent of the delay vector T, is

$$\mathcal{S}_0 := \{p \in \mathbf{P}_0 : \ F_p(z) \ne 0 \text{ in } \operatorname{Re}(z) \ge 0 \text{ for all } T \ge 0\} \tag{4.1}$$

when we write $F_p(z)$ instead of $F(z)$. This form of stability may be conceived as an extreme instance of system robustness. Our present aim is to find an explicit subset of $\mathcal{S}_0$.

$F(z)$ from (1.5) is, formally, a monic polynomial (with exponential polynomial coefficients), i.e., the coefficient of z^2 is a unity and so it is independent of T. This has the important consequence that each right half-plane $\mathrm{Re}(z) \geq x_0$, $x_0 \in \mathbf{R}$, contains only finitely many zeros of $F(z)$. One says that $F(z)$ is a *retarded* characteristic function. In the second part of this section, we shall consider the enlarged class of $F(z)$ given by

$$\begin{aligned} &F(z) := z^2C(z) + zA(z) + B(z), \\ &C(z) := \sum_{k=0}^{n} C_k e^{-zT_k}, \qquad C_k \in \mathbf{R}, \qquad C_0 := 1. \end{aligned} \tag{4.2}$$

The coefficients $A(z), B(z)$ are defined as in (1.5). $F(z)$ from (4.2) is said to be *neutral* if $|C_1| + |C_2| + \cdots + |C_n| > 0$. For our present purposes, it is appropriate to decompose F(z) from (4.2) into the delay-dependent part and the delay-independent one. So, henceforth we write for the $F(z)$ from the class (4.2),

$$\begin{aligned} &F(z) := A(z) + B(z), \\ &A(z) := z^2 + A_1 z + A_0 =: (z - a_1)(z - a_2), \qquad A_1, A_0 \in \mathbf{R}, \\ &B(z) := \sum_{k=1}^{n} B_k \cdot b_{k,p_k}(z) e^{-zT_k}, \qquad p_k \in \{0, 1, 2\}, \\ &b_{k,j}(z) := (z - b_k)^j, \qquad j = 0, 1, \qquad b_k \in \mathbf{R}, \\ &b_{k,2}(z) := (z - b_{k,1})(z - b_{k,2}). \end{aligned} \tag{4.3}$$

If $b_{k,1} \notin \mathbf{R}$ then $b_{k,2} := \bar{b}_{k,1}$, so that $b_{k,2}(z)$ is a real polynomial. One notices that the $b_{k,j}(z)$ are monic polynomials.

We focus our interest on the stability behaviour of $F(z)$ with varying delay vector T and also write $F_T(z) := F(z)$ when stressing the delay dependence of $F(z)$. It is known that an unstable $F_0(z)$ can become a stable $F_T(z)$ for $T > 0$ in some subsets of $\mathbf{T}$. When dealing with delay-independent stability as defined in (4.1), the polynomial $F_0(z)$ must be stable. For $B_k = 0$, $k = 1(1)n$, $A(z)$ must be stable. So, we restrict the class of $F(z)$ from (4.3) to the subclass from (4.3) with

$$A_1 > 0, \qquad A_0 > 0. \tag{4.4}$$

It is generally known (and also a consequence of Theorem 2.1) that (4.4) is equivalent to

$$\mathrm{Re}(a_1) < 0, \qquad \mathrm{Re}(a_2) < 0. \tag{4.5}$$

$A(z)$ is the characteristic function of a damped harmonic oscillator. Therefore, $F(z)$ from (4.3) may be viewed as a characteristic function of such a perturbed oscillator. So, the expectation is that $F(z)$ remains stable as long as the stability reserve of $A(z)$ is not yet exhausted by the perturbation $B(z)$. The magnitude B of the perturbation may be measured by a certain positive function of the moduli $|B_k|$, $k = 1(1)n$. We shall use notation from §2 in what follows.

THEOREM 4.1. *The function* $F(z)$ *from* (4.3), (4.4) *with* $p_k < 2$, $k = 1(1)n$, *is zero-free in* $\mathrm{Re}(z) \geq 0$ *for all nonnegative delay vectors* $T \geq 0$ *if*

$$|B_1| + |B_2| + \cdots + |B_n| < B, \tag{4.6}$$

where

$$\begin{aligned} B^2 &:= \min\{ M_k^2 : \ k = 1(1)n \}, \\ M_k^2 &:= \max\{ (1-p_k)M_0^2, p_k M_{k,1}^2 \}, \\ M_0^2 &:= \begin{cases} A_0^2, & 2A_0 \leq A_1^2, \\ A_1^2(A_0^2 - A_1^2/4), & 2A_0 > A_1^2, \end{cases} \\ M_{k,1}^2 &:= \begin{cases} A_0^2/b_k^2, & r_k \leq b_k^4, \\ \frac{(A_0+b_k^2-\sqrt{r_k})^2 + A_1^2(\sqrt{r_k}-b_k^2)}{\sqrt{r_k}}, & r_k > b_k^4, \end{cases} \\ r_k &:= (A_0 + b_k^2)^2 - A_1^2 b_k^2. \end{aligned} \tag{4.7}$$

Proof. We consider $F(z)$ from (4.3), (4.4) on the family of simple closed curves

$$C_R := -I_R \cup S_R, \qquad R > R_0 \tag{4.8}$$

for sufficiently large $R_0 > 0$. Rouché's theorem of the theory of functions tells us that $F(z)$ has as many zeros as $A(z)$ inside C_R if

$$|A(z)| > |B(z)| \tag{4.9}$$

on C_R. By hypothesis (4.5), $A(z) \neq 0$ in $\mathrm{Re}(z) \geq 0$ and so is $F(z)$ if (4.9) holds for all $R > R_0$. Now, (4.9) holds on S_R since $|A(z)| = O(|z^2|)$, $|B(z)| = O(|z|)$, $|z| \to \infty$, under the restriction that $p_k < 2$, $k = 1(1)n$. So, if (4.9) is fulfilled on the entire axis I then it holds on C_R for $R > R_0$. We shall show the validity of (4.9) along I under (4.6). For $z := iy$, $y \in \mathbf{R}$, one has

$$\begin{aligned} |A(z)|^2 &= (A_0 - y^2)^2 + A_1^2 y^2, \\ |B|^2 &\leq M^2 \max\{ (y^2 + b_k^2)^{p_k} : \ k = 1(1)n \}, \\ M &:= |B_1| + |B_2| + \cdots + |B_n|. \end{aligned} \tag{4.10}$$

Therefore, if

$$(A_0 - y^2)^2 + A_1^2 y^2 > M^2 \max\{ (y^2 + b_k^2)^{p_k} : \ k = 1(1)n \} \tag{4.11}$$

then (4.9) holds on I. The right-hand side of (4.11) becomes arbitrarily small for y in any compact set when M is made small enough. So, there are M^2 solving (4.11). In order to solve (4.11) for M^2, we write it in system form as

$$(A_0 - y^2)^2 + A_1^2 y^2 > M^2 (y^2 + b_k^2)^{p_k}, \qquad k = 1(1)n. \tag{4.12}$$

M^2 solves the system (4.12) if all system equations are solved. Consider the kth system equation of (4.12). If $p_k = 0$ then the system equation reduces to

$$(A_0 - y^2)^2 + A_1^2 y^2 > M^2, \tag{4.13}$$

so that all

$$M^2 < M_0^2 := \begin{cases} A_0^2, & 2A_0 \le A_1^2, \\ A_1^2(A_0^2 - A_1^2/4), & 2A_0 \ge A_1^2 \end{cases} \tag{4.14}$$

solve (4.13). In the remaining case $p_k = 1$, we put (4.12) in the shape

$$\begin{aligned} &\frac{(A_0 - y^2)^2 + A_1^2 y^2}{y^2 + b_k^2} > M^2, \\ &y^2 + b_k^2 + c_k + \frac{r_k}{y^2 + b_k^2} > M^2, \\ &\qquad r_k := (A_0 + b_k^2)^2 - A_1^2 b_k^2. \end{aligned} \tag{4.15}$$

The expression for $c_k \in \mathbf{R}$ is not of interest for us. The left-hand sides of the inequalities in (4.15) attain their common minimum either at $y^2 := 0$ or $y^2 := \sqrt{r_k} - b_k^2$. The latter takes place only for $r_k \ge b_k^4$. Inserting the location of the minimum in the first line of (4.15) yields the solution set

$$M^2 < M_{k,1}^2 := \begin{cases} A_0^2/b_k^2, & r_k < b_k^4, \\ \frac{(A_0 - \sqrt{r_k} + b_k^2)^2 + A_1^2(\sqrt{r_k} - b_k^2)}{\sqrt{r_k}}, & r_k \ge b_k^4. \end{cases} \tag{4.16}$$

We may unify both cases $p_k = 0, 1$ by defining the solution set of the kth system equation as

$$M^2 < M_k^2 := \max\{\, (1 - p_k) M_0^2, p_k M_{k,1}^2 \,\}. \tag{4.17}$$

Finally, the full system (4.12) is solved by

$$M^2 < B^2 := \min\{\, M_k^2 : k = 1(1)n \,\}, \tag{4.18}$$

and this is the assertion of the theorem. □

For moderate n, it poses no problem to reduce estimation losses by determining better bounds B numerically.

We turn to the neutral case. Let $a, \bar{a}$ be a pair of complex conjugate, not purely imaginary points if $a \notin \mathbf{R}$ and a pair of unrelated real points, different from the origin, if $a \in \mathbf{R}$. The same notation applies for b. We consider for $z := iy$, $y \in \mathbf{R}$,

$$F(y) := \left| \frac{(z - a)(z - \bar{a})}{(z - b)(z - \bar{b})} \right|^2 \tag{4.19}$$

and need an explicit representation for

$$M^2(a, \bar{a}, b, \bar{b}) := \inf\{\, F(y) : \; y \in \mathbf{R} \,\}. \tag{4.20}$$

With the notation

$$A_1 := \frac{a^2 + \bar{a}^2}{2}, \quad B_1 := \frac{b^2 + \bar{b}^2}{2}, \quad R_a^2 := a^2 \bar{a}^2, \quad R_b^2 := b^2 \bar{b}^2, \tag{4.21}$$

the function $F(y)$ is written as a function of $Y := y^2$ as

$$F(Y) := \frac{Y^2 + 2A_1 Y + R_a^2}{Y^2 + 2B_1 Y + R_b^2}. \tag{4.22}$$

LEMMA 4.2. *The quantity* M^2, *defined in* (4.20), *with the notation* (4.21) *has the value*

$$M^2(a,\bar{a},b,\bar{b}) := \begin{cases} 1, & B_1 \le A_1, \quad R_b \le R_a, \\ 1/r, & B_1 \le rA_1, \quad R_b > R_a, \\ m_1^2, & rA_1 \le B_1 \le A_1, \quad R_b > R_a, \\ m_2^2, & (\max\{A_1, rA_1\} < B_1) \text{ or } (B_1 > A_1,\ R_b < R_a), \end{cases}$$

$$r := \frac{R_b^2}{R_a^2},$$

$$m_j^2 := \frac{A_1 + Y_j}{B_1 + Y_j}, \quad j = 1,2, \tag{4.23}$$

$$Y_j := Y_0 + (-1)^j \cdot \sqrt{D + Y_0^2},$$

$$Y_0 := \frac{R_b^2 - R_a^2}{2A_1 - 2B_1},$$

$$D := \frac{B_1 R_a^2 - A_1 R_b^2}{B_1 - A_1}.$$

A proof of the foregoing lemma can be found in the appendix. It is elementary but rather tedious.

THEOREM 4.3. $F(z)$ *from* (4.3), (4.4) *is zero-free in* $\mathrm{Re}(z) \ge 0$ *for all nonnegative delay vectors* $T \ge 0$ *if*

$$|B_1| + |B_2| + \cdots + |B_n| < B, \tag{4.24}$$

where

$$B^2 := \min\{ M_k^2 : \ k = 1(1)n \},$$

$$M_k^2 := \begin{cases} M_0^2, & p_k = 0, \\ M_{k,1}^2, & p_k = 1, \\ M_{k,2}^2, & p_k = 2. \end{cases} \tag{4.25}$$

The quantities M_0^2, $M_{k,1}^2$ *are to be taken from* (4.7) *and*

$$M_{k,2}^2 := M^2(a_1, a_2, b_{k,1}, b_{k,2}), \tag{4.26}$$

formed with $M(a,\bar{a},b,\bar{b})$ *from Lemma* 4.2.

Proof. In the retarded case, i.e., all $p_k < 2$, the proof is that of Theorem 4.1. So, it suffices to add here only the reasoning for the indices k in the sum $B(z)$ with $p_k = 2$. If k is such an index, the corresponding system equation of the system (4.12) with $z := iy$, $y \in \mathbf{R}$, is written as

$$\begin{aligned} |(z-a_1)(z-a_2)| &> M|(z-b_{k,1})(z-b_{k,2})|, \\ \left|\frac{(z-a_1)(z-a_2)}{(z-b_{k,1})(z-b_{k,2})}\right|^2 &> M^2, \\ M^2(a_1,a_2,b_{k,1},b_{k,2}) &> M^2, \end{aligned} \tag{4.27}$$

which completes the proof. □

Observe that $B < 1$ when $F(z)$ is neutral.

Next we shall give yet another subset of $\mathcal{S}_0$. For our present purposes, it is convenient to write the retarded functions of the class (1.5) as

$$\begin{aligned} F_T(z) &:= A(z) + B(z), \qquad T := (T_1, T_2, \ldots, T_n) \geq 0, \\ A(z) &:= z^2 + Az + 1, \qquad A > 0, \\ B(z) &:= \sum_{k=1}^{n} (B_k z + C_k) e^{-zT_k}, \qquad B_k, C_k \in \mathbf{R}, \qquad T_k \in \mathbf{R}_+. \end{aligned} \tag{4.28}$$

There is no loss of generality in assuming $A(0) = 1$ in the class of stable $A(z)$ of degree two. In preparing the next theorem, we consider the two parameter family of real rational functions

$$f(y) := \frac{y + a}{y^2 + 2cy + 1}, \qquad a \geq 0, \qquad c > -1. \tag{4.29}$$

LEMMA 4.4. *The maximum*

$$M := \max\{ f(y) : \ y \geq 0 \} \tag{4.30}$$

formed with $f(y)$ *from* (4.29) *has the value*

$$M := M(a, c) := \begin{cases} \frac{1}{2\left(c - a + \sqrt{a^2 - 2ac + 1}\right)}, & 2ac < 1, \\ a, & 2ac \geq 1. \end{cases} \tag{4.31}$$

Proof. For $c > -1$, $f(y)$ possesses no poles in $y \geq 0$ and $f(y)$ is continuous there. As a real function, $f(y)$ has a pair of complex-conjugated critical points (i.e., zeros of $f'(y)$) which solve

$$\begin{aligned} (y + a)^2 &= a^2 - 2ac + 1 = (a - c)^2 + 1 - c^2, \\ y_j &:= -a + (-1)^j \cdot \sqrt{a^2 - 2ac + 1}, \qquad j = 1, 2. \end{aligned} \tag{4.32}$$

For $c \leq (a + 1/a)/2$, the pair $y = y_1, y_2$ is real and y_2 may not lie in $y \geq 0$. In terms of the parameters, the requirement $y_2 > 0$ means $1 > 2ac$. We observe that $f(y) > 0$ for $y > -a$ and $f(y) < 0$ for $y < -a$. The cases $y_2 \leq 0$ and $y_2 > 0$ shall be considered separately.

Case 1. $y_2 \leq 0$. In this case, $[0, +\infty)$ is free of critical points and $f(y)$ decreases in $[0, +\infty)$. This means

$$M := f(0) = a, \qquad 2ac \geq 1. \tag{4.33}$$

Case 2. $y_2 > 0$. In this case, $y = y_2$ necessarily locates a maximum on $[0, +\infty)$ and we have $M := f(y_2)$. After an elementary calculation we find

$$M := \frac{1}{2\left(c - a + \sqrt{a^2 - 2ac + 1}\right)}, \qquad 2ac < 1. \tag{4.34}$$

Unifying the definitions (4.33), (4.34) gives the expression (4.31). □

Later we will need the inverse function $A(m, c)$ of $M(a, c)$ with respect to the first argument a.

LEMMA 4.5. *The solution set of*

$$m > M(a,c), \qquad a \geq 0, \qquad c > -1 \tag{4.35}$$

with $M(a,c)$ *from* (4.31) *is the set*

$$0 \leq a < A(m,c), \qquad c > \frac{1}{2m} - 1, \tag{4.36}$$

where

$$\begin{aligned} A(m,c) &:= \begin{cases} m, & c \geq \frac{1}{2m}, \\ a(m,c), & \frac{1}{2m} - 1 \leq c \leq \frac{1}{2m}, \end{cases} \\ a(m,c) &:= m\left[1 - \left\{\frac{1}{2m} - c\right\}^2\right]. \end{aligned} \tag{4.37}$$

Proof. We consider inequality (4.35) in $2ac < 1$ and $2ac \geq 1$ separately. In the second case, (4.35) reads $m > a$ so that

$$\frac{1}{2c} \leq a < m, \qquad c \geq \frac{1}{2m} \tag{4.38}$$

in the part of the searched solution set lying in $2ac \geq 1$. In $2ac < 1$, we have

$$\begin{aligned} 2m\left[c - a + \sqrt{a^2 - 2ac + 1}\right] &> 1, \\ a &< m\left[1 - \left\{\frac{1}{2m} - c\right\}^2\right] =: a(c,m). \end{aligned} \tag{4.39}$$

The function $a(c,m)$ is nonnegative in the interval $1/(2m) - 1 \leq c \leq 1/(2m) + 1$ and $[1/(2m) - 1, 1/(2m)]$ is a subinterval thereof. So, the solution set in $2ac < 1$ is

$$0 \leq a < a(m,c), \qquad \frac{1}{2m} - 1 \leq c < \frac{1}{2m}. \tag{4.40}$$

The union of both sets (4.38) and (4.40) is (4.36) with (4.37), which completes the proof. □

The boundary curve $A(m,c)$ of the solution set is C^1 and consists of a half-line parallel to the c axis and a parabolic arc joining the points $(0, \frac{1}{2m} - 1)$ and $(m, \frac{1}{2m})$ in the (a,c) plane.

In the next proof we shall refer to the inequality for $a_k \geq 0$, $k = 1(1)n$,

$$\left(\sqrt{a_1} + \sqrt{a_2} + \cdots + \sqrt{a_n}\right)^2 \leq n(a_1 + \cdots + a_n). \tag{4.41}$$

This special case of the Schwartz inequality can be shown using the identity

$$(\sqrt{a} - \sqrt{b})^2 \equiv a + b - 2\sqrt{ab}. \tag{4.42}$$

While (4.41) becomes an equality for $a_k = a$ all equal, the estimation losses grow with n and the degree of straying of the points a_k. So, for $a_1 = a_2 = \cdots = a_{n-1} = 0$, $a_n = a$, (4.41) estimates $a \leq na$.

THEOREM 4.6. *$F_T(z)$ from (4.28) is zero-free in $\mathrm{Re}(z) \geq 0$ for all delay vectors $T \geq 0$ if*

$$C_1^2 + \cdots + C_n^2 < (B_1^2 + \cdots + B_n^2)\, A\left(\frac{1}{n \cdot (B_1^2 + \cdots + B_n^2)}, \frac{A^2}{2} - 1\right) \tag{4.43}$$

with $A(m, c)$ from (4.37).

Proof. We follow the argumentation in the proof of Theorem 4.1 up to formula (4.10). The result is that $F_T(z)$ is zero-free in $\mathrm{Re}(z) \geq 0$ if

$$|A(z)| > |B(z)| \tag{4.44}$$

holds on $\mathrm{Re}(z) = 0$. We set $z = iy$, $y \in \mathbf{R}$, and may rewrite (4.44) as

$$\begin{aligned} &1 > \max\{\, Q_T(y) : \; y \in \mathbf{R},\; T \geq 0 \,\} =: M, \\ &Q_T(y) := \left|\frac{B(z)}{A(z)}\right|^2. \end{aligned} \tag{4.45}$$

We compute M iteratively; first we maximize over $T \geq 0$ and then over $y \in \mathbf{R}$. The bound $M^2(y)$ in

$$|B(z)|^2 \leq (\sqrt{B_1^2 y^2 + C_1^2} + \cdots + \sqrt{B_n^2 y^2 + C_n^2})^2 =: M^2(y) \tag{4.46}$$

is best possible if the components of T are independent. We have

$$\begin{aligned} M^2(y) &:= B^2 y^2 + C^2 + 2 \sum_{1 \leq j < k \leq n} \sqrt{(B_j^2 y^2 + C_j^2)(B_k^2 + C_k^2)} \geq B^2 y^2 + C^2, \\ C^2 &:= C_1^2 + \cdots + C_n^2, \qquad B := B_1^2 + \cdots + B_n^2. \end{aligned} \tag{4.47}$$

We use (4.41) in (4.46) and afterwards apply Lemma 4.4. In the notation $Y := y^2$, we have, for $B^2 > 0$,

$$\begin{aligned} \left|\frac{B(z)}{A(z)}\right|^2 &\leq \frac{n(B^2 Y + C^2)}{Y^2 + 2\left(\frac{A^2}{2} - 1\right) Y + 1}, \\ &\leq n B^2 M\left[\frac{C^2}{B^2}, \frac{A^2}{2} - 1\right]. \end{aligned} \tag{4.48}$$

Using (4.48) in (4.45) yields the sufficient condition for delay-independent stability

$$\frac{1}{n} > B^2 M\left[\frac{C^2}{B^2}, \frac{A^2}{2} - 1\right]. \tag{4.49}$$

The right-hand side of the first line in (4.48) vanishes for $C^2 + B^2 \to 0$. So, (4.49) specifies a nonempty set in parameter space for all $n \in \mathbf{N}$ and all $A > 0$. In order to obtain an explicit representation of this set, we solve (4.49) for the first argument C^2/B^2 of M and have, by Lemma 4.5,

$$C^2 < B^2 A\left[\frac{1}{nB^2}, \frac{A^2}{2} - 1\right]. \tag{4.50}$$

Do not confound the function $A(m,c)$ and the constant A from (4.28) in (4.50)! The expanded form of (4.50) is (4.43). □

5. Discussion. Given a representation for N, the number of zeros of $F(z)$ in $\mathrm{Re}(z) \geq 0$, $N = 0$ is a stability criterion. Let $F(z) := F_p(z)$ be a member of a parametric family $p \in \mathbf{P}$ of functions $F(z)$. Then $N = 0$ solved for p is a stability chart. From this point of view, the selection of an appropriate representation is decisive for what it is built upon. In §2, N was represented in terms of the pair of real functions $U(y) := \mathrm{Re}[F(iy)]$, $V(y) := \mathrm{Im}[F(iy)]$. As already mentioned, Theorem 2.1 is a specialization of a theorem that is valid for a class of meromorphic functions with polynomial growth for $|z| \to \infty$ in $\mathrm{Re}(z) \geq 0$. The asymmetric use of $U(y)$ and $V(y)$ can be avoided. This possibility and normalizations are based on the simple fact that $F(z)$ and $G(z) := F(z) \cdot M(z)$, where $M(z) \neq 0$ in $\mathrm{Re}(z) \geq 0$ have the same number of zeros in $\mathrm{Re}(z) \geq 0$.

The representation of N in terms of $U(y)$, $V(y)$ was selected because numerical computation of N was the main goal. In this representation, the determination of N comprises the following steps (we assume $F(z)$ adheres to the class (2.3)):

(5.1)
1. Determine a bound Y for all positive zeros of $U(y)$.
2. Determine all sign changes y_k, $k = 1(1)m$, of $U(y)$ in $[0, Y]$.
3. Compute the sign sequence σ with elements $\sigma_k := \mathrm{sgn}[V(y_k)]$, $k = 1(1)m$.
4. Compute N via formula (2.18).

Procedure (5.1) is straightforward and the numerical effort seems reasonable for the yield. In the presence of rounding errors, only zeros of $F(z)$ in a close neighbourhood of the origin can lead to false signs σ_k. Critical signs in that respect can be detected and handled differently. (In our counterexample we took the analytical approach.)

Other representations of N are known. Hermite [7] represented N as signature of a certain quadratic form. Hurwitz [8] used this representation as point of departure for his famous determinantal criterion for a polynomial $F(z)$ to be zero-free in $\mathrm{Re}(z) \geq 0$.

In the Cauchy index theorem (see Gantmacher [6, pp. 524–528], Marden [11, p. 169], or Cauchy [4]) N is defined in terms of N_-, N_+. The latter and the index are defined as $I_{-\infty}^{+\infty} Q := N_+ - N_-$, where

$$
\begin{aligned}
N_- &:= \sum_{Q_k<0} 1, \qquad N_+ := \sum_{Q_k>0} 1, \\
Q(y) &:= \frac{V(y)}{U(y)} := \sum_k \frac{Q_k}{y - y_k} + R(y),
\end{aligned}
\tag{5.2}
$$

where y_k runs through the real simple roots of $U(y)$. For simplicity, all real roots are assumed to be simple. For the general case, cf. Gantmacher [6]. Consequently, the remainder term $R(y)$ comprises all further members in the partial fraction expansion of the quotient $Q(y)$ of the polynomials $U(y), V(y)$. In §2, we have shown how N can be represented in terms of the pair $\{U(y), V(y)\}$ instead of the quotient $Q(y)$. Our division-free formulation avoids unnecessary numerical computations.

A class of stability criteria is formulated in terms of geometrical or topological features of the curve $F(iy)$, $y \in \mathbf{R}$. The Nyquist criterion (cf. [12]) reposes on the winding number. One has to plot (a sufficiently long arc of) $F(iy)$ in $\mathbf{C}$ and then the winding number is to be read off. The Hermite–Biehler theorem allows one to decide

about the stability of the generating polynomial on the basis of the relative positions of the sequence of real zeros of the pair $\{U(y), V(y)\}$. This theorem has been extended to the largest class of entire transcendental functions; see Postnikov [15] or Lewin [10, Chap. 9]. The class of exponential polynomials $F(z)$ with real exponential points $T_1 < T_2 < \cdots < T_n$, positive defect $D := (T_1 + T_n)/2 > 0$, and real polynomial coefficients $A_k(z)$ belongs (see [10, Ex. 1, p. 324,] to this class;

$$F_T(z) := \sum_{k=1}^{n} A_k(z) e^{-zT_k}. \tag{5.3}$$

The Hermite–Biehler property is one of the stability characterizing features which could be carried over from polynomials to entire transcendental functions. To mention further versions of stability criteria is beyond the scope of this section.

We consider the retarded exponential polynomial with nonnegative exponential points T_k and real polynomial coefficients $A_k \in \mathbf{R}[z]$;

$$\begin{aligned} &F_T(z) := \sum_{k=0}^{n} A_k(z) e^{-zT_k}, \qquad 0 =: T_0 < T_1 < \cdots < T_n, \qquad n > 1, \\ &A_k(z) := A_k z^{d_k} + \cdots + A_{k,0}, \qquad k = 0(1)n, \qquad d_0 > d_k \qquad A_k \neq 0. \end{aligned} \tag{5.4}$$

All but the leading coefficients are doubly indexed. The zeros of $F_T(z)$ of large modulus $r := |z|$ are accessible to asymptotical analysis. In what follows it suffices to consider only the asymptotical $|z| \to \infty$ relevant part $F_T^0(z)$ of $F_T(z)$. We write the coefficients as

$$A_k(z) := A_k z^{d_k}[\,1 + o(1)\,], \qquad |z| \to \infty \tag{5.5}$$

and define

$$F_T^0(z) := \sum_{k=0}^{n} A_k z^{d_k} e^{-zT_k}. \tag{5.6}$$

We look for an appropriate enclosure for all zeros of $F_T^0(z)$. Let z be a zero of $F_T^0(z)$; then, with $z := x + iy$, $x, y \in \mathbf{R}$, we have

$$\begin{aligned} -A_0 z^{d_0} &= \sum_{k=1}^{n} A_k z^{d_k} e^{-zT_k}, \\ |A_0| &\leq (A - |A_0|) \max\left\{ r^{-(d_0 - d_k)} e^{-xT_k} : \; k = 1(1)n \right\}, \\ B_0 := \frac{A - |A_0|}{|A_0|} &\leq \max\left\{ r^{-(d_0 - d_k)} e^{-xT_k} : \; k = 1(1)n \right\}, \\ A &:= |A_0| + \cdots + |A_n|. \end{aligned} \tag{5.7}$$

In order to solve the last inequality of (5.7) for $|y|$, we rewrite it in system form as

$$\begin{aligned} r^2 &\leq C_k^2 e^{-2xU_k}, \qquad k = 1(1)n, \\ U_k &:= \frac{T_k}{d_0 - d_k}, \\ C_k &:= B_0^{-\frac{1}{d_0 - d_k}} \end{aligned} \tag{5.8}$$

The solution of the kth system equation is

$$y^2 \le C_k^2 e^{-2xU_k} - x^2 =: y_k^2(x). \tag{5.9}$$

If $y_k^2(x) < 0$ then no zeros of the kth system equation are possible on the vertical line $\mathrm{Re}(z) = x$. From (5.9), for the solution of the full system (5.8), follows

$$|y| \le \max\{\, y_k(x) :\ k = 1(1)n \,\} =: y_0(x), \qquad x \le x_0 := \max\{\, x_k :\ k = 1(1)n \,\} \tag{5.10}$$

with the understanding that $F_T^0(z)$ is zero-free on $\mathrm{Re}(z) = x$ if $y_0^2(x) < 0$ The x_k are the unique positive roots of $y_k^2(x)$. We call the set

$$\mathbf{L} := \{\, z \in \mathbf{C} :\ \mathrm{Re}(z) \le x,\ \mathrm{Im}(z) = y_0(x),\ x \le x_0 \,\} \tag{5.11}$$

a *left logarithmic half-plane.* It is bounded by a C^0 spline $y_0(x)$ along which $\mathrm{Re}(z)$ decreases logarithmically with $|\mathrm{Im}(z)|$ for $|\mathrm{Im}(z)| \to \infty$. The enclosure of all zeros of $F_T^0(z)$ in a logarithmic left half-plane $\mathbf{L}$ implies that this is also true for $F_T(z)$ itself (for another $\mathbf{L}$). For $T \to 0$, $\mathbf{L}$ becomes doubly connected. The rightmost component of $\mathbf{L}$ becomes a disk for $T = 0$ centered about the origin $z = 0$, and the second component lies in the left half-plane $\mathrm{Re}(z) \le x_0$, where $x_0 \to -\infty$ for $T \to 0$.

Appendix. We shall consider the function

$$F(Y) := \frac{Y^2 + 2A_1Y + R_a^2}{Y^2 + 2B_1Y + R_b^2}, \qquad R_a^2 \ge A_1^2 > 0, \qquad R_b^2 \ge B_1^2 > 0 \tag{A.1}$$

and determine the infimum

$$M^2 := M^2(A_1, B_1, R_a, R_b) := \inf\{\, F(Y) :\ Y \ge 0 \,\}. \tag{A.2}$$

Since M^2 is a C^0 spline on its four-dimensional domain of definition, no simple representation can be expected. An explicit piecewise representation in the following lemma corresponds to the spline character of M^2.

LEMMA A.1. *The function M^2 defined by* (A.2) *and* (A.1) *allows the representation*

$$M^2 := \begin{cases} 1, & B_1 \le A_1, \quad R_b \le R_a, \\ 1/r, & B_1 \le rA_1, \quad R_b > R_a, \\ m_1^2, & rA_1 \le B_1 \le A_1, \quad R_b > R_a, \\ m_2^2, & (\max\{A_1, rA_1\} < B_1) \ \text{or} \ (B_1 > A_1,\ R_b < R_a), \end{cases}$$

$$r := \frac{R_b^2}{R_a^2},$$

$$m_j^2 := \frac{A_1 + Y_j}{B_1 + Y_j}, \quad j = 1, 2, \tag{A.3}$$

$$Y_j := Y_0 + (-1)^j \sqrt{D + Y_0^2},$$

$$Y_0 := \frac{R_b^2 - R_a^2}{2A_1 - 2B_1},$$

$$D := \frac{B_1 R_a^2 - A_1 R_b^2}{B_1 - A_1}.$$

Proof. In the first step we compile properties of $F(Y)$ used in a later case distinction. In our approach, it is appropriate to consider $F(Y)$ on the whole real line $\mathbf{R}$ and not only for $Y \geq 0$ as demanded by (A.2). Asymptotics for $Y \to 0$ and $|Y| \to \infty$ yield

$$\begin{aligned} F(Y) :&= R_a^2 R_b^{-2} \cdot \frac{1 + \frac{2A_1 Y}{R_a^2} + O(Y^2)}{1 + \frac{2B_1 Y}{R_b^2} + O(Y^2)}, \qquad R_b \neq 0, \qquad Y \to 0, \\ &= R_a^2 R_b^{-2} \cdot \left[1 + 2\left(\frac{A_1}{R_a^2} - \frac{B_1}{R_b^2}\right) Y + O(Y^2)\right], \\ &= \frac{1 + \frac{2A_1}{Y} + O(Y^{-2})}{1 + \frac{2B_1}{Y} + O(Y^{-2})}, \qquad |Y| \to \infty, \\ &= 1 + 2(A_1 - B_1)Y + O(Y^{-2}). \end{aligned} \tag{A.4}$$

Hence, with $r := R_b^2 / R_a^2$,

$$\begin{aligned} \operatorname{sgn}[F'(0)] &= \operatorname{sgn}(rA_1 - B_1), \\ \operatorname{sgn}[F(Y) - 1] &= \operatorname{sgn}(Y) \operatorname{sgn}(A_1 - B_1), \qquad |Y| \to \infty. \end{aligned} \tag{A.5}$$

Now, we determine the solutions Y_0 of $F(Y) = 1$. These are the solutions of

$$2(A_1 - B_1)Y = R_b^2 - R_a^2. \tag{A.6}$$

Clearly, for $A_1 \neq B_1$ there is exactly one solution $Y = Y_0$,

$$Y_0 := \frac{R_b^2 - R_a^2}{2A_1 - 2B_1}. \tag{A.7}$$

For $A_1 = B_1$ and $R_a \neq R_b$ there is no solution of (A.6) or, more precisely,

$$\begin{aligned} F(Y) > 1, \qquad A_1 = B_1, \qquad R_a > R_b, \\ F(Y) < 1, \qquad A_1 = B_1, \qquad R_a < R_b. \end{aligned} \tag{A.8}$$

In the degenerate case, $A_1 = B_1$, $R_a = R_b$, we have $F(Y) \equiv 1$. In the nondegenerate case, the extrema of $F(Y)$ are located at the solutions Y of

$$\frac{F'}{F} = \frac{2(Y + A_1)}{Y^2 + 2A_1 Y + R_a^2} - \frac{2(Y + B_1)}{Y^2 + 2B_1 Y + R_b^2} = 0. \tag{A.9}$$

Evidently, for $A_1 = B_1$, there is exactly one solution $Y = -A_1$ which, in the first case of (A.8), locates the maximum and in the second case the minimum $F(-A_1) = (R_a^2 - A_1^2)/(R_b^2 - A_1^2)$. For $A_1 \neq B_1$, (A.9) may be rewritten as

$$\begin{aligned} (Y - Y_0)^2 &= D + Y_0^2 =: \Delta, \\ Y = Y_j &:= Y_0 + (-1)^j \sqrt{D + Y_0^2}, \qquad j = 1, 2, \\ Y_1 &:= -A_1, \qquad A_1 = B_1, \\ D &:= \frac{B_1 R_a^2 - A_1 R_b^2}{B_1 - A_1}, \\ &= \frac{R_a^2 (B_1 - rA_1)}{B_1 - A_1}. \end{aligned} \tag{A.10}$$

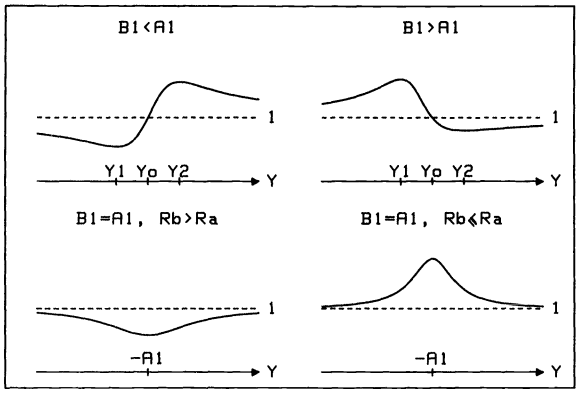


FIG. A.1. *The main cases* $B_1 < A_1$, $B_1 = A_1$, $B_1 > A_1$ *of* $F(Y) := \frac{Y^2+2A_1Y+R_a^2}{Y^2+2B_1Y+R_b^2}$; *dashed asymptote* $F(Y) :\equiv 1$.

The definition of Y_1 for $A_1 = B_1$ is possible because the singularity in the general expression for Y_1 is removable. According to (A.10), the roots Y_1, Y_2 lie symmetrically about $Y = Y_0$ irrespective of the sign of the discriminant Δ. In fact, $\Delta > 0$. In order to see this, we observe that $F'/F = 2(A_1 - B_1) \neq 0$ for $Y = Y_0$ in the case considered. This means that $F(Y)-1$ has a sign change at $Y = Y_0$. Rolle's theorem, applied to the intervals $(-\infty, Y_0)$ and $(Y_0, +\infty)$ tells us that each of them contains exactly one root of $F'(Y)$. So, there is no double root of $F'(Y)$, whence $\Delta > 0$. For $B_1 > A_1$, $F(Y_1)$ is the maximum of $F(Y)$ and $F(Y_2)$ is the minimum. For $B_1 < A_1$, the roles of Y_1, Y_2 are reversed; $F(Y_1)$ is the minimum and $F(Y_2)$ is the maximum. If Y is a root of F'/F then

$$F(Y) := \frac{Y^2 + 2A_1Y + R_a^2}{Y^2 + 2B_1Y + R_b^2} = \frac{A_1 + Y}{B_1 + Y} =: m^2(Y). \tag{A.11}$$

We shall use the notation $m_j^2 := m^2(Y_j)$, $j = 1, 2$.

The above knowledge enables us to determine M^2. In higher-dimensional parameter spaces there is a virtual propensity to overlook some regions. The countermeasure used here involves painstaking bookkeeping (not enjoining the readership) in complete case distinctions.

We shall treat A_1, R_a as independent parameters and B_1, R_b as the dependent ones. At the first level, we discern the three cases $A_1 < B_1$, $A_1 = B_1$, $A_1 > B_1$. In the first and last cases, subcases defined by the position of the origin $Y = 0$ relative to the three points $Y_1 < Y_0 < Y_2$ will be considered. It is helpful to consult Fig. A.1 during the subsequent case distinction.

Case 1. $B_1 > A_1$. Recall that $Y = Y_1$ locates the maximum and $Y = Y_2$ locates the minimum in the present case. We shall consider the subcases $Y_2 \leq 0$, $Y_2 > 0$, which is a complete case distinction. These conditions are expressed in terms of the four parameters of $F(Y)$.

Subcase 1.1. $B_1 > A_1$, $Y_2 \leq 0$. Since $F(Y)$ increases in $Y \geq 0$ for $Y_2 \leq 0$, we have $M^2 := F(0)$. Furthermore, $Y_2 \leq 0$ implies $Y_0 < 0$ and $D \leq 0$. According to (A.7), (A.10), this means

$$M^2 := 1/r, \qquad A_1 < B_1 \leq rA_1, \qquad R_b > R_a. \tag{A.12}$$

Clearly, (A.12) is vacuous for $A_1 \leq 0$.

Subcase 1.2. $B_1 > A_1$, $Y_2 > 0$. The minimum of $F(Y)$ is located to the right of $Y = 0$ so that $M^2 := F(Y_2)$. We shall consider the subcases $F'(0) < 0$ and $F(0) > 1$. The second case defines the interval $(-\infty, Y_0)$ and the first defines (Y_1, Y_2); so both cases overlap on (Y_1, Y_0). The second subcase entails $R_b < R_a$.

Subcase 1.2.1. $B_1 > A_1$, $Y_1 < 0 < Y_2$. We have $rA_1 < B_1$ for $Y_1 < 0 \leq Y_2$ so that

$$M^2 := m_2^2, \qquad \max\{A_1, rA_1\} < B_1. \tag{A.13}$$

Subcase 1.2.2. $B_1 > A_1$, $R_b < R_a$. Here, trivially,

$$M^2 := m_2^2, \qquad B_1 > A_1, \qquad R_b < R_a. \tag{A.14}$$

Case 2. $B_1 = A_1$. We partition the present case into the subcases $R_b \leq R_a$ and $R_b > R_a$.

Subcase 2.1. $A_1 = B_1$, $R_b \leq R_a$. Uniting the degenerate case $F(Y) \equiv 1$ and the first case of (A.8) gives

$$M^2 := 1, \qquad A_1 = B_1, \qquad R_b \leq R_a. \tag{A.15}$$

Subcase 2.2. $A_1 = B_1$, $R_b > R_a$. $F(Y)$ has a nondecreasing and a nonincreasing part. We therefore consider the tertiary subcases $F'(0) \geq 0$ and $F'(0) \leq 0$.

Subcase 2.2.1. $A_1 = B_1$, $R_b > R_a$, $F'(0) \geq 0$. From $M^2 := F(0)$ in the nondecreasing part follows

$$M^2 := 1/r, \qquad A_1 = B_1 \leq rA_1, \qquad R_b > R_a. \tag{A.16}$$

Subcase 2.2.2. $A_1 = B_1$, $R_b > R_a$, $F'(0) \leq 0$. Since the minimum of $F(Y)$ lies to the right of $Y = 0$, we obtain $M^2 := F(Y_1)$, where $Y_1 := -A_1$,

$$M^2 := m_1^2, \qquad A_1 = B_1 \geq rA_1, \qquad R_b > R_a. \tag{A.17}$$

Case 3. $B_1 < A_1$. The monotonicity behaviour of $F(Y)$ directs us to consider the subcases $Y_0 \leq 0$, $Y_1 < 0 < Y_0$, $Y_1 \geq 0$.

Subcase 3.1. $B_1 < A_1$, $Y_0 \leq 0$. From $F(Y) \geq 1$ in $Y \geq 0$ follows

$$M^2 := 1, \qquad B_1 < A_1, \qquad R_b \leq R_a. \tag{A.18}$$

Subcase 3.2. $B_1 < A_1$, $Y_1 < 0 < Y_0$. In the present case, $F(Y) < 1$ increases in $0 \leq Y < Y_0$ so that $M^2 := F(0)$. The condition $Y_1 < 0 < Y_0$ implies $R_b > R_a$ and $D > 0$. This leads us to

$$M^2 := 1/r, \qquad B_1 < \min\{A_1, rA_1\}, \qquad R_b > R_a. \tag{A.19}$$

Subcase 3.3. $B_1 < A_1$, $Y_1 \geq 0$. The minimum of $F(Y)$ is assumed at $Y = Y_1$ and lies at $Y = 0$ or to the right of the origin $Y = 0$. Therefore $M^2 := F(Y_1)$. From $Y_1 \geq 0$ follows $Y_0 > 0$ and $D \leq 0$. This means

$$M^2 := m_1^2, \qquad rA_1 \leq B_1 < A_1, \qquad R_b > R_a. \tag{A.20}$$

We now unite the sets in parameter space for which M^2 has the same values and obtain the representation in (A.3). Verify that (A.12), (A.16), (A.19) allow the representation of the second case in (A.3). □

It should be mentioned that the representation of M^2 by (A.3) is not based on a partition of the parameter space spanned by (A_1, B_1, R_a, R_b) but on a covering. A less concise representation of M^2 can be found in [2].

Acknowledgment. In a private communication to the author in 1990, Dr. G. Stépán from the Department of Mechanics of the Technical University of Budapest expressed doubts about the correctness of the stability criterion invoked in [5]. This was the starting point for this paper.

REFERENCES

[1] L. V. AHLFORS, *Complex Analysis*, McGraw-Hill, New York, 1966.
[2] F. G. BOESE, *Delay-independent stability of a special sequence of neutral difference-differential equations with one delay*, J. Differential Equations, 90 (1991), pp. 397–407
[3] ———, *A Necessary and Sufficient Stability Criterion for a Class of Meromorphic Functions of One Complex Variable*, manuscript.
[4] A. L. CAUCHY, *Calcul des indices des fonctions*, J. de l'Ecole Polytechnique, 15 (1837), pp. 176–229.
[5] H. I. FREEDMAN AND V. S. H. RAO, *Stability criteria for a system involving two time delays*, SIAM J. Appl. Math., 46 (1986), pp. 552–660
[6] F. R. GANTMACHER, *Matrizentheorie*, Springer-Verlag, Berlin, 1986.
[7] C. HERMITE, *Sur l'indice des fractions rationelles*, Bull. Soc. Math. France, 7 (1879), pp. 128–131.
[8] A. HURWITZ, *Über die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Teilen besitzt*, in Mathematische Werke von Adolf Hurwitz, Birkhäuser, Basel, 1933, Bd. 2, pp. 533–545.
[9] A. M. KRALL, *Stability Techniques for Continuous Linear Systems*, Gordon and Breach, New York, 1967.
[10] B. J. LEWIN, *Nullstellenverteilung ganzer Funktionen*, Akademie-Verlag, Berlin, 1962.
[11] M. MARDEN, *The geometry of polynomials*, in Amer. Math. Soc. Surveys III, American Mathematical Society, Providence, RI, 1966.
[12] H. NYQUIST, *Regeneration theory*, Bell System Technical Journal, 11 (1932), pp. 126–147.
[13] N. OBRESCHKOFF, *Verteilung und Berechnung der Nullstellen reeller Polynome*, Deutscher Verlag der Wissenschaften, Berlin, 1963.
[14] B. PORTER, *Stability Criteria for Linear Dynamical Systems*, Oliver and Boyd, Edinburgh, 1967.
[15] M. M. POSTNIKOV, *Stable Polynomials*, Nauka, Moscow, 1981. (In Russian.)
[16] G. STÉPÁN, *Retarded Dynamical Systems: Stability and Characteristic Functions*, Longman Scientific and Technical, Harlow, UK, 1989.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1331–1341, September 1995

013

ON THE INVERSION OF THE LAPLACE TRANSFORM OF C_0 SEMIGROUPS AND ITS APPLICATIONS*

PENG-FEI YAO†

Abstract. The main results are as follows: (a) The inversion of the Laplace transform of C_0 semigroup representation in a Hilbert space H is generalized to all $x \in H$. (b) A full characterization is given in terms of the resolvent $R(\lambda; A)$ of the infinitesimal generator A of a C_0 semigroup $T(t)$ in a Hilbert space, which assures the continuity of $T(t)$ for $t > 0$ in the uniform operator topology.

Key words. C_0 semigroup, Laplace transform, Fourier transform

AMS subject classification. 47D06

1. Introduction. First, we recall some basic concepts concerning abstract integration. Let H be a Hilbert space (the inner product and the induced norm in H are denoted by $\langle ., . \rangle$ and $\| \cdot \|$, respectively), $\mathbb{R} = (-\infty, \infty)$, $I \subset \mathbb{R}$. Suppose that $f : \mathbb{R} \times \mathbb{R} \longrightarrow H$ is an H-valued function. We say that the H-valued Bochner integral $\int_{\mathbb{R}} f(t,s)ds$ converges uniformly in $t \in I$ if the Lebesgue integral $\int_{\mathbb{R}} \|f(t,s)\|ds$ converges uniformly in $t \in I$, and the H-valued improper Riemann integral $\int_{\mathbb{R}} f(t,s)ds$ converges uniformly in $t \in I$ if, for each $\epsilon > 0$, there is $N > 0$ independent of $t \in I$ such that, when $b > a > N$ or $a < b < -N$,

$$\left\| \int_a^b f(t,s)ds \right\| < \varepsilon \quad \forall t \in I.$$

It is trivially checked that the H-valued improper integral $\int_{\mathbb{R}} f(t,s)ds$ exists if the H-valued Bochner integral $\int_{\mathbb{R}} f(t,s)ds$ exists, and the two integrals are equal when $f : \mathbb{R} \times \mathbb{R} \longrightarrow H$ is a continuous H-valued function.

Now we consider the inversion of the Laplace transform representing a C_0 semigroup. A related classical theorem (see [1, Chap. 1, Thm. 1.7], for example) may be described as follows.

Let A be the infinitesimal generator of a C_0 semigroup $T(t)$ on H. Denote by $D(A)$ the domain of A and by $R(\lambda; A)$ the resolvent of A. Let

$$\alpha_0 > \lim_{t \to +\infty} \frac{\log \|T(t)\|}{t}.$$

If $x \in D(A^2)$ then

$$T(t)x = \frac{e^{\alpha_0 t}}{2\pi} \int_{\mathbb{R}} e^{it\tau} R(\alpha_0 + i\tau; A)x d\tau. \tag{1.1}$$

The right-hand side integral of (1.1) is a Bochner integral and it converges uniformly in $t \in [\delta, \frac{1}{\delta}](\delta > 0)$.

In this paper, we shall prove that formula (1.1) holds for all $x \in H$ with the proviso that the right-hand side integral of (1.1) is an H-valued improper Riemann integral. This is expressed in the following theorem.

* Received by the editors November 9, 1992; accepted for publication (in revised form) January 27, 1994. This research was supported by the National Natural Science Foundation of China.
† Institute of Systems Science, Academia Sinica, Beijing 100080, People's Republic of China.

THEOREM 1.1. *Let A be the infinitesimal generator of a C_0 semigroup $T(t)$ on a Hilbert space H and*

$$\alpha_0 > \lim_{t\to+\infty} \frac{\log \|T(t)\|}{t}.$$

Then

$$T(t)x = \frac{n!e^{\alpha_0 t}}{2\pi t^n} \int_{\mathbb{R}} e^{it\tau} R^{n+1}(\alpha_0 + i\tau; A)x\,d\tau, \quad t > 0, x \in H, n = 0, 1, 2, \ldots. \tag{1.2}$$

The integrals $\int_{\mathbb{R}} e^{it\tau} R^{n+1}(\alpha_0 + i\tau; A)x\,d\tau$ are H-valued improper Riemann integrals. Moreover, the integral $\int_{\mathbb{R}} e^{it\tau} R(\alpha_0 + i\tau; A)x\,d\tau$ converges uniformly in $t \in [\delta, \frac{1}{\delta}](\delta > 0)$ and the integrals $\int_{\mathbb{R}} e^{it\tau} R^{n+1}(\alpha_0 + i\tau; A)x\,d\tau$ converge uniformly in $t \in \mathbb{R}$ when $n \geq 1$.

Finally, we consider characteristic conditions of a C_0 semigroup $T(t)$, which assure the continuity of $T(t)$ for $t > 0$ in the uniform operator topology. Pazy has twice pointed out (see [1, pp. 50 and 256]) that "so far, there are no known necessary and sufficient conditions, in terms of A or the resolvent $R(\lambda; A)$, which assure the continuity for $t > 0$ of $T(t)$ in the uniform operator topology." As an application of Theorem 1.1, we have the following theorem.

THEOREM 1.2. *Let A be the infinitesimal generator of a C_0 semigroup $T(t)$ on a Hilbert space H and*

$$\alpha_0 > \lim_{t\to+\infty} \frac{\log \|T(t)\|}{t}.$$

Then $T(t)$ is continuous for $t > 0$ in the uniform operator topology if and only if

$$\sup_{x\in H, \|x\|=1} \int_{|\tau|\geq a} \|R(\alpha_0 + i\tau; A)x\|^2 d\tau \to 0 \quad (as \quad a \to +\infty). \tag{1.3}$$

Furthermore, if $T(t)$ is continuous for $t > 0$ in the uniform operator topology, then

$$T(t) = \frac{n!e^{\alpha_0 t}}{2\pi t^n} \int_{\mathbb{R}} e^{it\tau} R^{n+1}(\alpha_0 + i\tau, A)d\tau, \quad n = 0, 1, 2, \ldots. \tag{1.4}$$

The integrals $\int_{\mathbb{R}} e^{it\tau} R^{n+1}(\alpha_0 + i\tau; A)d\tau$ are operator-valued improper Riemann integrals. The integral $\int_{\mathbb{R}} e^{it\tau} R(\alpha_0 + i\tau; A)d\tau$ converges uniformly in $t \in [\delta, \frac{1}{\delta}](\delta > 0)$ and the integrals $\int_{\mathbb{R}} e^{it\tau} R^{n+1}(\alpha_0 + i\tau; A)d\tau$ converge uniformly in $t \in \mathbb{R}$ when $n \geq 1$.

From the above theorem, together with [1, Chap. 2, Thm. 2.3], Corollary 1.3 immediately follows.

COROLLARY 1.3. *Let A be the infinitesimal generator of a C_0 semigroup $T(t)$ on a Hilbert space H and*

$$\alpha_0 > \lim_{t\to+\infty} \frac{\log \|T(t)\|}{t}.$$

Then $T(t)$ is a compact semigroup if and only if

(a) *$R(\lambda; A)$ is compact for $\lambda \in \rho(A)$, the resolvent set of A;*

(b)

$$\sup_{x\in H, \|x\|=1} \int_{|\tau|\geq a} \|R(\alpha_0 + i\tau; A)x\|^2 d\tau \to 0 \quad (as \quad a \to +\infty).$$

Moreover, if $T(t)$ is a compact semigroup, formulae (1.4) hold.

2. Proof of Theorem 1.1. To begin with, we build several lemmas.

For any $p \geq 1$, set

$$L^P(\mathbb{R}; H) = \{f \mid f : \mathbb{R} \longrightarrow H \text{ such that } \int_{\mathbb{R}} \|f(t)\|^P dt < +\infty\}.$$

Then $L^2(\mathbb{R}; H)$ is a Hilbert space with the inner product $(f, g)_{L^2(\mathbb{R};H)} = \int_{\mathbb{R}} \langle f(t), g(t)\rangle dt$ and the induced norm $\|\cdot\|_{L^2(\mathbb{R};H)}$, while $L^1(\mathbb{R}; H)$ is a Banach space.

Let $f \in L^2(\mathbb{R}; H) \bigcap L^1(\mathbb{R}; H)$. Denote by

$$F(f)(\tau) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-it\tau} f(t) dt \tag{2.1}$$

the Fourier transform of f. We have the following lemma.

LEMMA 2.1. *The mapping F, defined by* (2.1), *can be extended to the whole* $L^2(\mathbb{R}; H)$ *to become a unitary operator from* $L^2(\mathbb{R}; H)$ *to itself. This extension is still denoted by* F.

Proof. By emulating the proofs of [2, Chap. I, Thms. 2.1 and 2.3], we get the above lemma. A point for attention is the replacement of the convolution of functions with $\int_{\mathbb{R}} \langle f(t+s), g(s)\rangle ds$, where $f, g \in L^2(\mathbb{R}; H)$.

LEMMA 2.2. *Let A be the infinitesimal generator of a* C_0 *semigroup* $T(t)$ *and*

$$\alpha_0 > \lim_{t \to +\infty} \frac{\log \|T(t)\|}{t}.$$

For simplicity, denote $R_{\alpha_0}(\tau) = R(\alpha_0 + i\tau; A)$ *and* F^{-1} = *inversion of* F. *Then*

(a) $R^{n+1}_{\alpha_0}(\cdot) \in L^2(\mathbb{R}; H)$ *for all* $x \in H$, $n = 0, 1, 2, \ldots$;

(b) $F^{-1}(R^{n+1}_{\alpha_0} x)(t) = (t^n/n!) F^{-1}(R_{\alpha_0} x)(t), t \in \mathbb{R}, x \in H, n = 0, 1, 2, \ldots$;

(c)

$$T(t)x = \frac{e^{\alpha_0 t}}{\sqrt{2\pi}} F^{-1}(R_{\alpha_0} x)(t) \quad a.e. \quad t > 0, \quad x \in H. \tag{2.2}$$

Proof. Take α_1 such that

$$\alpha_0 > \alpha_1 > \lim_{t \to +\infty} \frac{\log \|T(t)\|}{t}.$$

Then there is a constant $M > 0$ such that

$$\|T(t)\| \leq M e^{\alpha_1 t} \quad \forall \quad t \geq 0. \tag{2.3}$$

(a) For every $x \in H$ set

$$G_{\alpha_0}(t) = \begin{cases} e^{-\alpha_0 t} T(t)x, & t \geq 0, \\ 0 \quad t < 0. \end{cases} \tag{2.4}$$

Since $\lim_{t \to 0+} T(t)x = x$ together with (2.3), we have

$$G_{\alpha_0}(\cdot)x \in L^2(\mathbb{R}; H). \tag{2.5}$$

By the inversion of the Laplace transform, we have

$$\begin{aligned} R_{\alpha_0}(\tau)x &= \int_0^{+\infty} e^{-(\alpha_0+i\tau)t}T(t)xdt \\ &= \int_0^{+\infty} e^{-it\tau}e^{-\alpha_0 t}T(t)xdt \\ &= \sqrt{2\pi}\frac{1}{\sqrt{2\pi}}\int_0^{+\infty} e^{-it\tau}G_{\alpha_0}(t)xdt \\ &= \sqrt{2\pi}F(G_{\alpha_0}x)(\tau), \quad \tau \in \mathbb{R}. \end{aligned} \tag{2.6}$$

From Lemma 2.1, F is a unitary mapping from $L^2(\mathbb{R};H)$ to itself. From (2.6), we have

$$R_{\alpha_0}(\cdot)x \in L^2(\mathbb{R};H)$$

or

$$\int_{\mathbb{R}} \|R_{\alpha_0}(\tau)x\|^2 d\tau < +\infty. \tag{2.7}$$

In addition, from the Hille–Yosida theorem there is a constant $K > 0$ such that

$$\|R_{\alpha_0}(\tau)\| \le K \quad \forall \tau \in \mathbb{R}. \tag{2.8}$$

Thus, from (2.7) and (2.8) we obtain

$$\int_{\mathbb{R}} \|R_{\alpha_0}^{n+1}(\tau)x\|^2 d\tau \le K^{2n}\int_{\mathbb{R}} \|R_{\alpha_0}(\tau)x\|^2 d\tau < +\infty$$

or, equivalently,

$$R_{\alpha_0}^{n+1}(\cdot)x \in L^2(\mathbb{R};H), \quad x \in H, \quad n = 0,1,2,\ldots$$

(b) From the well-known resolvent identity it follows that

$$R_{\alpha_0}^{(n)}(\tau)x = (-i)^n n! R_{\alpha_0}^{n+1}(\tau)x, \quad x \in H, \quad n = 0,1,2,\ldots.$$

Using the properties of the Fourier transform, we have

$$\begin{aligned} F^{-1}(R_{\alpha_0}^{n+1}x) &= \frac{i^n}{n!}F^{-1}(R_{\alpha_0}^{(n)}x)(t) \\ &= \frac{t^n}{n!}F^{-1}(R_{\alpha_0}x)(t), \quad t \in \mathbb{R}, \quad x \in H, \quad n = 0,1,2,\ldots. \end{aligned}$$

(c) From (2.4) together with (2.6) we obtain

$$T(t)x = e^{\alpha_0 t}G_{\alpha_0}(t)x = \frac{e^{\alpha_0 t}}{\sqrt{2\pi}}F^{-1}(R_{\alpha_0}x)(t) \quad \text{a.e.} \quad t > 0, \quad x \in H.$$

LEMMA 2.3. *Denote by $\mathcal{L}(H)$ the Banach space of all bounded linear operators from H to itself with the uniform operator topology. Let f and $g : \mathbb{R} \longrightarrow \mathcal{L}(H)$ be continuous in the uniform operator topology. Moreover, suppose that f satisfies*

$$\sup_{t\in\mathbb{R}} \|f(t)\| < +\infty$$

and g satisfies the following inequalities:

$$\int_{\mathbb{R}} \|g(\tau)x\|^2 d\tau < +\infty \quad \forall x \in H$$

and

$$\int_{\mathbb{R}} \|g^*(\tau)x\|^2 d\tau < +\infty \quad \forall x \in H,$$

where $g^*(\tau)$ is the adjoint of $g(\tau)$. Then

(a) *there are constants $\beta > 0$ and $\mu > 0$ such that*

$$\int_{\mathbb{R}} \|g(\tau)x\|^2 d\tau \leq \beta^2 \|x\|^2$$

and

$$\int_{\mathbb{R}} \|g^*(\tau)x\|^2 d\tau \leq \mu^2 \|x\|^2 \quad \forall x \in H; \tag{2.9}$$

(b) *for any $a < b$ we have*

$$\left\| \int_a^b g(\tau)f(\tau)g(\tau)x d\tau \right\| \leq \mu \sup_{t \in \mathbb{R}} \|f(t)\| \left(\int_a^b \|g(\tau)x\|^2 d\tau \right)^{\frac{1}{2}} \tag{2.10}$$

for all $x \in H$.

Proof. (a) Define a linear operator $B : H \longrightarrow L^2(\mathbb{R}; H)$ by

$$(Bx)(\tau) = g(\tau)x \quad \forall x \in H \quad \forall \tau \in \mathbb{R}.$$

It is easily checked that B is a closed mapping with $D(B) = H$. Thus the closed graph theorem leads us to the conclusion that B is a bounded linear operator from H to $L^2(\mathbb{R}; H)$. Hence

$$\int_{\mathbb{R}} \|g(\tau)x\|^2 d\tau \leq \|B\|^2 \|x\|^2 \quad \forall x \in H.$$

Then, taking $\beta = \|B\|$ yields the desired inequality. A similar argument shows that there is a constant $\mu > 0$ such that the inequality of (2.9) holds.

(b) From (a), for any $x, y \in H$ we have

$$\begin{aligned}
&\left| \left\langle \int_a^b g(\tau)f(\tau)g(\tau)x d\tau, y \right\rangle \right| \\
&= \left| \int_a^b \langle f(\tau)g(\tau)x, g^*(\tau)y \rangle d\tau \right| \\
&\leq \int_a^b \|f(\tau)\| \|g(\tau)x\| \|g^*(\tau)y\| d\tau \\
&\leq \sup_{t \in \mathbb{R}} \|f(t)\| \left(\int_a^b \|g(\tau)x\|^2 d\tau \right)^{\frac{1}{2}} \left(\int_a^b \|g^*(\tau)y\|^2 d\tau \right)^{\frac{1}{2}}.
\end{aligned}$$

The proof is then complete.

Proof of Theorem 1.1. Let $n \geq 1$ and take $K > 0$ such that

$$\|R_{\alpha_0}(\tau)\| \leq K \quad \forall \tau \in I\!R. \tag{2.11}$$

Note that the adjoint semigroup $T^*(t)$ of $T(t)$ is still a C_0 semigroup which is generated by the adjoint operator A^* of A, and

$$\lim_{t\to+\infty} \frac{\log \|T(t)\|}{t} = \lim_{t\to+\infty} \frac{\log \|T^*(t)\|}{t}.$$

Then, applying Lemma 2.2 (a) to $T(t)$ and $T^*(t)$, respectively, we have

$$\int_{I\!R} \|R_{\alpha_0}(\tau)x\|^2 d\tau < +\infty \quad \forall x \in H \tag{2.12}$$

and

$$\begin{aligned} \int_{I\!R} \|R^*_{\alpha_0}(\tau)x\|^2 d\tau &= \int_{I\!R} \|R(\alpha_0 - i\tau; A^*)x\|^2 d\tau \\ &= \int_{I\!R} \|R(\alpha_0 + i\tau; A^*)x\|^2 d\tau < +\infty \quad \forall x \in H. \end{aligned} \tag{2.13}$$

Set $g(\tau) = R_{\alpha_0}(\tau)$ and $f_t(\tau) = e^{it\tau} R^{n-1}_{\alpha_0}(\tau)$ for $t, \tau \in I\!R$. From (2.11)–(2.13) it is clear that g and f_t satisfy the conditions of Lemma 2.3. Thus we have that, for arbitrary $a < b$,

$$\begin{aligned} \left\| \int_a^b e^{it\tau} R^{n+1}_{\alpha_0}(\tau)x d\tau \right\| &= \left\| \int_a^b g(\tau) f_t(\tau) g(\tau)x d\tau \right\| \\ &\leq \mu \sup_{s\in I\!R} \|f_t(s)\| \left(\int_a^b \|g(\tau)x\|^2 d\tau \right)^{\frac{1}{2}} \\ &\leq \mu K^{n-1} \left(\int_a^b \|R_{\alpha_0}(\tau)x\|^2 \right)^{\frac{1}{2}}, \quad t \in I\!R, \quad x \in H, \end{aligned} \tag{2.14}$$

where μ is a constant independent of a, b. Since the Lebesgue integral

$$\int_{I\!R} \|R_{\alpha_0}(\tau)x\|^2 d\tau$$

exists, the improper Riemann integral

$$P_n(t)x = \int_{I\!R} e^{it\tau} R^{n+1}_{\alpha_0}(\tau)x d\tau, \quad n \geq 1 \tag{2.15}$$

converges uniformly in $t \in I\!R$. Thus $P_n(\cdot)x : I\!R \longrightarrow H$ is an H-valued continuous function.

For each $a > 0$ set

$$Q(\tau, a)x = \begin{cases} R^{n+1}_{\alpha_0}(\tau)x, & |\tau| \leq a, \\ 0, & |\tau| > a, \end{cases} \qquad n \geq 1.$$

It is easily checked that $Q_{\alpha_a}(\cdot)x$ converges to $R_{\alpha_0}^{n+1}(\cdot)x$ as $a \longrightarrow +\infty$ in the topology of $L^2(\mathbb{R}; H)$. Since F^{-1} is a bounded linear operator from $L^2(\mathbb{R}; H)$ to itself, we have

$$F^{-1}(\mathbb{R}_{\alpha_0}^{n+1}x)(t) = \lim_{a\to+\infty} F^{-1}(Q(\cdot,a)x)(t) = \lim_{a\to+\infty} \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} e^{it\tau} R_{\alpha_0}^{n+1}(\tau)x\,d\tau$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{it\tau} R_{\alpha_0}^{n+1}(\tau)x\,d\tau \quad \text{a.e.} \quad t \in \mathbb{R}, \quad x \in H \tag{2.16}$$

(the above limits are taken in $L^2(\mathbb{R}; H)$). From Lemma 2.2 (b) and (c) we have

$$T(t)x = \frac{e^{\alpha_0 t}}{\sqrt{2\pi}} F^{-1}(R_{\alpha_0}x)(t) = \frac{n!e^{\alpha_0 t}}{\sqrt{2\pi t^n}} F^{-1}(R_{\alpha_0}^{n+1}x)(t)$$

$$= \frac{n!e^{\alpha_0 t}}{2\pi t^n} \int_{\mathbb{R}} e^{it\tau} R_{\alpha_0}^{n+1}(\tau)(t)x\,d\tau \quad \text{a.e.} \quad t > 0, \quad x \in H, \quad n = 1, 2, 3, \ldots. \tag{2.17}$$

From (2.15) the two sides of equality (2.16) are continuous H-valued functions for $t > 0$ in the topology of H. Therefore we conclude that equality (2.16) holds for all $t > 0$.

Finally, we prove (1.2) for the case $n = 0$.

Setting $n = 1$ and $t = 0$ in (2.15), we see that the improper integral $\int_{\mathbb{R}} R_{\alpha_0}^2(\tau)x\,d\tau$ is convergent in the topology of H. Since

$$R_{\alpha_0}(b)x = R_{\alpha_0}(0)x - i\int_0^b R_{\alpha_0}^2(\tau)x\,d\tau,$$

the limits

$$\lim_{b\to+\infty} R_{\alpha_0}(b)x \quad \text{and} \quad \lim_{b\to-\infty} R_{\alpha_0}(b)x$$

exist in the topology of H. Using this fact together with (2.12), we have

$$\lim_{|\tau|\to+\infty} R_{\alpha_0}(\tau)x = 0 \quad \forall x \in H. \tag{2.18}$$

Since

$$\frac{d}{d\tau}(e^{it\tau}R_{\alpha_0}(\tau)x) = ite^{it\tau}R_{\alpha_0}(\tau)x - ie^{it\tau}R_{\alpha_0}^2(\tau)x$$

for any $a, b \in \mathbb{R}$, we have

$$e^{itb}R_{\alpha_0}(b)x - e^{ita}R_{\alpha_0}(a)x = it\int_a^b e^{it\tau}R_{\alpha_0}(\tau)x\,d\tau - i\int_a^b e^{it\tau}R_{\alpha_0}^2(\tau)x\,d\tau$$

or, equivalently,

$$\int_a^b e^{it\tau}R_{\alpha_0}(\tau)x\,d\tau = \frac{1}{t}\int_a^b e^{it\tau}R_{\alpha_0}^2(\tau)x\,d\tau - \frac{i}{t}[e^{itb}R_{\alpha_0}(b)x - e^{ita}R_{\alpha_0}(a)x], \tag{2.19}$$

$$t > 0, \quad x \in H, \quad a, b \in \mathbb{R}.$$

Then from (2.15), (2.17), and (2.18) it follows that the improper Riemann integral

$$\int_{\mathbb{R}} e^{it\tau}R_{\alpha_0}(\tau)x\,d\tau$$

converges uniformly in $t \in [\delta, 1/\delta](\delta > 0)$ in the topology of H. Finally, as in (2.16), we get the desired formula for the case $n = 0$.

3. Proof of Theorem 1.2. The following lemma will play a key role in proving Theorem 1.2.

LEMMA 3.1. *Let A be the infinitesimal generator of a C_0 semigroup $T(t)$ and*

$$\alpha_0 > \alpha_1 > \lim_{t\to+\infty} \frac{\log\|T(t)\|}{t}.$$

Then there is a constant $M > 0$ satisfying

$$\|T(t)\| \le Me^{\alpha_1 t} \quad \forall t \ge 0. \tag{3.1}$$

Let $G_{\alpha_0}(t)x$ be defined by (2.4), $x \in H$, and denote $T_1(t) = e^{-\alpha_0 t}T(t)$, $t \ge 0$. Set

$$h_x(t) = \int_{\mathbb{R}} \langle G_{\alpha_0}(t+s)x, G_{\alpha_0}(s)x\rangle ds \quad \forall t \in \mathbb{R} \quad \forall x \in H$$

and

$$Z(t) = \int_0^{+\infty} \|T_1(t+s) - T_1(s)\| ds \quad \forall t \ge 0.$$

If $T(t)$ is continuous in the uniform operator topology for $t > 0$, then we have

(a) $0 \le Z(t) \le M(M+1)/(\alpha_0 - \alpha_1) \quad \forall t \ge 0$;

(b) $\lim_{t\to 0^+} Z(t) = 0$;

(c) $|h_x(t) - h_x(0)| \le Z(|t|)M\|x\|^2 \quad \forall t \in \mathbb{R}\ \forall x \in H$;

(d) *the existence of a constant $l > 0$ such that*

$$|h_x(t)| \le l\|x\|^2 \quad \forall t \in \mathbb{R} \quad \forall x \in H.$$

Proof. (a) From (3.1) we have

$$\begin{aligned} 0 \le Z(t) &= \int_0^{+\infty} \|[T_1(t) - I]T_1(s)\| ds \\ &\le M(M+1)\int_0^{+\infty} e^{-(\alpha_0-\alpha_1)s} ds \\ &= \frac{M(M+1)}{\alpha_0 - \alpha_1} \quad \forall t \ge 0. \end{aligned}$$

(b) The continuity of $T_1(t)$ in the uniform operator topology for $t > 0$ is a direct result of the continuity of $T(t)$. Hence

$$\|T_1(t+s) - T_1(t)\| \longrightarrow 0 \quad \text{as} \quad t \longrightarrow 0^+ \quad \forall s > 0$$

and

$$\|T_1(t+s) - T_1(s)\| \le (M+1)\|T_1(s)\| \quad \forall s \ge 0.$$

Using the dominated convergence theorem, we get (b).

(c) For $t < 0$ we have

$$\begin{aligned} h_x(t) &= \int_{-t}^{+\infty} \langle G_{\alpha_0}(t+s)x, G_{\alpha_0}(s)x\rangle ds \\ &= \int_{-t}^{+\infty} \langle T_1(t+s)x, T_1(-t)T_1(t+s)x\rangle ds \\ &= \int_0^{+\infty} \langle T_1(w)x, T_1(-t+w)x\rangle dw. \end{aligned}$$

Hence

$$\begin{aligned}|h_x(t) - h_x(0)| &= \left| \int_0^{+\infty} \langle T_1(w)x, [T_1(-t+w) - T_1(w)]x \rangle dw \right| \\ &\leq \int_0^{+\infty} \|T_1(-t+w) - T_1(w)\| dw M \|x\|^2 \\ &= Z(|t|)M\|x\|^2 \quad \forall x \in H.\end{aligned}$$

Similarly, we can obtain the desired conclusion for the case $t > 0$.

(d) From (3.1) we have

$$\int_{\mathbb{R}} \|G_{\alpha_0}(t)x\|^2 dt < +\infty \quad \forall x \in H.$$

By Lemma 2.3(a) there is a constant $\beta > 0$ such that

$$\begin{aligned}\int_0^{+\infty} \|T_1(s)x\|^2 ds &= \int_{\mathbb{R}} \|G_{\alpha_0}(t)x\|^2 dt \\ &\leq \beta^2 \|x\|^2 \quad \forall x \in H.\end{aligned}$$

Thus we have

$$\begin{aligned}|h_x(t)| &\leq Z(|t|)M\|x\|^2 + |h_x(0)| \\ &\leq \left[\frac{M^2(M+1)}{\alpha_0 - \alpha_1} + \beta^2 \right] \|x\|^2 \quad \forall x \in H.\end{aligned}$$

Therefore, to obtain the conclusion of (d), we only need to take $l = M^2(M+1)/(\alpha_0 - \alpha_1) + \beta^2$. The proof is then finished.

Proof of Theorem 1.2. We first prove sufficiency. Take $n = 1$ in formula (1.2); we have

$$T(t)x = \frac{e^{\alpha_0 t}}{2\pi t} P_1(t)x \quad \forall t > 0 \quad \forall x \in H,$$

where

$$P_1(t)x = \int_{\mathbb{R}} e^{it\tau} R^2_{\alpha_0}(\tau)x d\tau \quad \forall t \in \mathbb{R} \quad \forall x \in H. \tag{3.2}$$

Since the improper Riemann integral

$$\int_{\mathbb{R}} e^{it\tau} R^2_{\alpha_0}(\tau)x d\tau$$

converges uniformly in $t \in \mathbb{R}$ in the topology of H, the linear operator $P_1(t) : H \longrightarrow H$ defined by (3.2) is bounded for each $t \in \mathbb{R}$. For the sufficiency it remains to prove that $P_1(t)$ is continuous in the uniform operator topology for $t \in \mathbb{R}$.

For $t, t_0 \in \mathbb{R}$, and $x \in H$, from the inequality (2.14), for any fixed $a > 0$, we have

$$\begin{aligned}\|P_1(t)x - P_1(t_0)x\| &= \left\| \int_{\mathbb{R}} (e^{it\tau} - e^{it_0\tau}) R^2_{\alpha_0}(\tau)x d\tau \right\| \\ &\leq \int_{-a}^{a} |e^{it\tau} - e^{it_0\tau}| \|R^2_{\alpha_0}(\tau)\| d\tau \|x\| + \mu \left(\int_{|\tau| \geq a} \|R_{\alpha_0}(\tau)x\|^2 d\tau \right)^{\frac{1}{2}}\end{aligned}$$

or

$$\|P_1(t)-P_1(t_0)\| \le \int_{-a}^{a} |e^{it\tau}-e^{it_0\tau}| \|R^2_{\alpha_0}(\tau)\| d\tau + \mu \sup_{x\in H,\|x\|=1} \left(\int_{|\tau|\ge a} \|R_{\alpha_0}(\tau)x\|^2 d\tau \right)^{\frac{1}{2}}.$$

Using the dominated convergence theorem together with hypothesis (1.3), the desired continuity of $P_1(t)$ is deduced.

Now we prove necessity. We use the notation of Lemma 3.1. Denote by $\hat{h}_x(\tau)$ the Fourier transform of the complex function $h_x(t)$, where $x \in H$. By elementary calculations we obtain

$$\hat{h}_x(\tau) = \sqrt{2\pi}\|R_{\alpha_0}(\tau)x\|^2 \quad \forall \tau \in \mathbb{R} \quad \forall x \in H. \tag{3.3}$$

Thus, to prove the necessity it suffices only to prove

$$\sup_{x\in H,\|x\|=1} \int_{|\tau|\ge a} \hat{h}_x(\tau)d\tau \to 0 \quad (\text{as} \quad a \to +\infty). \tag{3.4}$$

For each $R > 0$ we have

$$\begin{aligned} \int_{-R}^{R} \left(1 - \frac{|\tau|}{R}\right) e^{-it\tau} d\tau &= \frac{1}{R} \int_0^R dw \int_{-w}^{w} e^{it\tau} d\tau \\ &= \frac{4\sin^2 \frac{Rt}{2}}{Rt^2} \quad \forall t \in \mathbb{R}, \quad t \neq 0. \end{aligned}$$

Hence for any fixed $R > 0$,

$$\begin{aligned} \int_{-R}^{R} \left(1 - \frac{|\tau|}{R}\right) \hat{h}_x(\tau)d\tau &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h_x(t) \left(\int_{-R}^{R} e^{-it\tau} \left(1 - \frac{|\tau|}{R}\right) d\tau \right) dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h_x(t) \frac{4\sin^2 \frac{Rt}{2}}{Rt^2} dt \\ &= \frac{2}{\sqrt{2\pi}} \int_{\mathbb{R}} h_x\left(\frac{2w}{R}\right) \frac{\sin^2 w}{w^2} dw \quad \forall x \in H. \end{aligned} \tag{3.5}$$

From (3.5) and the nonnegativity of $\hat{h}(\tau)$, by using Lemma 3.1 (a), (c), and (d) it follows that, for any $r > a > 0$,

$$\begin{aligned} 0 &\le \frac{1}{2} \int_{r\ge|\tau|\ge a} \hat{h}_x(\tau)d\tau \\ &\le \int_{r\ge|\tau|\ge a} \left(1 - \frac{|\tau|}{2r}\right) \hat{h}_x(\tau)d\tau \\ &\le \int_{2r\ge|\tau|\ge a} \left(1 - \frac{|\tau|}{2r}\right) \hat{h}_x(\tau)d\tau + \int_{-a}^{a} \left(\frac{1}{a} - \frac{1}{2r}\right) |\tau| \hat{h}_x(\tau)d\tau \\ &= \int_{-2r}^{2r} \left(1 - \frac{|\tau|}{2r}\right) \hat{h}_x(\tau)d\tau - \int_{-a}^{a} \left(1 - \frac{|\tau|}{a}\right) \hat{h}_x(\tau)d\tau \\ &= \sqrt{\frac{2}{\pi}} \int_{\mathbb{R}} \left[h_x\left(\frac{w}{r}\right) - h_x\left(\frac{2w}{a}\right) \right] \frac{\sin^2 w}{w^2} dw \end{aligned}$$

$$\leq \sqrt{\frac{2}{\pi}} \int_{-b}^{b} \left| h_x\left(\frac{w}{r}\right) - h_x(0) \right| dw + \sqrt{\frac{2}{\pi}} \int_{-b}^{b} \left| h_x\left(\frac{2w}{a}\right) - h_x(0) \right| dw$$
$$+2l\sqrt{\frac{2}{\pi}} \int_{|w|\geq b} \frac{\sin^2 w}{w^2} dw \|x\|^2$$
$$\leq M\sqrt{\frac{2}{\pi}} \int_{-b}^{b} \left[Z\left(\left|\frac{w}{r}\right|\right) + Z\left(\left|\frac{2w}{a}\right|\right) \right] dw\|x\|^2 + 2l\sqrt{\frac{2}{\pi}} \int_{|w|\geq b} \frac{\sin^2 w}{w^2} dw\|x\|^2$$

for any $b > 0$, where the following inequality is used:

$$\frac{\sin^2 w}{w^2} \leq 1 \quad \forall w \in I\!R, \quad w \neq 0.$$

Hence

$$0 \leq \frac{1}{2} \sup_{x\in H, \|x\|=1} \int_{r\geq|\tau|\geq a} \hat{h}_x(\tau) d\tau$$
$$\leq M\sqrt{\frac{2}{\pi}} \int_{-b}^{b} \left[Z\left(\left|\frac{w}{r}\right|\right) + Z\left(\left|\frac{2w}{a}\right|\right) \right] dw + 2l\sqrt{\frac{2}{\pi}} \int_{|w|\geq b} \frac{\sin^2 w}{w^2} dw$$

for any $b > 0$ and $r > a > 0$. From Lemma 3.1 and the dominated convergence theorem, together with the fact that the Lebesgue integral

$$\int_{I\!R} \frac{\sin^2 w}{w^2} dw$$

converges, we get assertion (3.4).

Finally, let $T(t)$ be continuous in the uniform operator topology for $t > 0$. From the above proof assertion (3.4) holds. Using this fact together with the inequality (2.14), we conclude that the operator-valued improper Riemann integral

$$\int_{I\!R} e^{it\tau} R^{n+1}_{\alpha_0}(\tau) d\tau, \quad n = 1, 2, 3, \ldots,$$

converges uniformly in $t \in I\!R$ in the uniform operator topology. Following the proof of Theorem 1.1, we know that the operator-valued improper Riemann integral

$$\int_{I\!R} e^{it\tau} R_{\alpha_0}(\tau) d\tau$$

converges uniformly in $t \in [\delta, \frac{1}{\delta}] (\delta > 0)$ in the uniform operator topology. Thus formulae (1.4) hold. The proof is then complete.

Acknowledgment. I would like to express my thanks to Professor De Xing Feng for his valuable advice.

REFERENCES

[1] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[2] E. M. STEIN AND G. WEISS, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, New Jersey, 1971.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1342–1360, September 1995

014

HIGHER SINGULARITIES AND FORCED SECONDARY BIFURCATION*

BERNHARD RUF†

Abstract. Singularity theory is used to study the solution structure of nonlinear differential equations. First, a characterization of the *fold, cusp, swallowtail,* and *butterfly singularities* is given in terms of derivatives of the zero eigenvalue of the linearization of the corresponding nonlinear operator. As an application a forced elliptic boundary value problem with cubic nonlinearity is considered: $-\Delta u - \lambda u + u^3 = f$ in $\Omega, \partial u/\partial n = 0$ on $\partial\Omega$. It has been previously shown that, for $\lambda_1 = 0 < \lambda < \lambda_2/7$, the corresponding nonlinear operator has only fold and cusp singularities, and for $0 < \lambda < \lambda_2/12$ the above equation has at most three solutions. Here we show that for $\Omega = (0,1)$ higher singularities develop for λ above and near $\lambda_2 \cdot 2/\pi^2$. These singularities can be identified as swallowtail and butterfly singularities. This can be interpreted as the appearance of a forced secondary bifurcation, since for certain forcing terms f there now exist at least five solutions.

Key words. singularity theory, bifurcation

AMS subject classifications. Primary, 58F14; Secondary, 58C27

1. Introduction.

In recent years methods of singularity theory have been successfully applied to the study of bifurcation phenomena in nonlinear differential equations. We refer, e.g., to Cafagna and Donati [5], [6], McKean and Scovel [15], and to the books of Golubitsky and Guillemin [11], Golubitsky and Schaeffer [12], and Golubitsky, Stewart, and Schaeffer [13]. Consider a given nonlinear differential equation as an operator equation between suitable Banach spaces E and F, say

$$\Phi(u) = f, \quad \text{where } u \in E \quad \text{and } f \in F \text{ is given;} \tag{1}$$

here it is assumed that the nonlinear mapping $\Phi : E \to F$ is of class $C^k(E,F)$ and Fredholm index zero (see §2 below). We are interested in obtaining information on the number of solutions of (1) in dependence on the forcing term f. Hence, f can be viewed as a bifurcation parameter. In order to obtain a characterization of possible bifurcation points, one determines, in a first step, the set S of *singular points* of Φ, that is, the points $u \in E$ in which Φ is not invertible: $S = \{u \in E; \exists v \in E\backslash\{0\}$ such that $\Phi'(u)[v] = 0\}$. Then, to obtain a local characterization of the mapping Φ in a singular point, one needs to determine its singularity type (in the Banach space analogue of the classifications of Whitney, Thom, and Arnold). For the simplest singularities—*fold* and *cusp*—the local structure of smooth Fredholm mappings has recently been characterized by Berger and Church [3] and Berger, Church, and Timourian [4] (see also Lazzeri and Micheletti [14], Cafagna and Donati [6]).

In this paper we will give the classification of the two subsequent singularities of so-called Morin type, the *swallowtail* and *butterfly* singularities.

As an application we will consider the elliptic boundary value problem

$$\begin{aligned} -\Delta u - \lambda u + u^3 &= h \quad \text{in } \Omega, \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2}$$

*Received by the editors February 3, 1993; accepted for publication (in revised form) January 5, 1994.

†Dipartimento di Matematica, Università degli Studi di Milano, Via Saldini 50, 20133 Milano, Italy.

Here $\Omega \subset R^n$ is a bounded and smooth domain, $\lambda \in R$ is a parameter, and $h \in C^{0,\alpha}(\Omega)$ with $\alpha \in (0,1)$ fixed is a given forcing term. We then consider $\Phi : E \to F$ with $E = \{u \in C^{2,\alpha}(\Omega); \frac{\partial u}{\partial n}|_{\partial\Omega} = 0\}$ and $F = C^{0,\alpha}(\Omega)$, and $\Phi(u) = -\Delta u - \lambda u + u^3$. It is well known that the mapping $-\Delta u + u$ is an isomorphism between E and F. In the case of $\Omega = (0,1)$ we may choose $E = \{u \in C^2(0,1); u'(0) = u'(1) = 0\}$ and $F = C(0,1)$. The mapping Φ is in $C^k(E,F)$ for every $k \geq 0$; in fact we have $\Phi'(u)[v] = -\Delta v + 3u^2 v$, $\Phi''(u)[v,w] = 6uvw$, $\Phi^{(3)}(u)[v,w,z] = 6vwz$, and $\Phi^{(k)}(u) = 0$ for $k \geq 4$.

Equation (2) has been studied in detail in [18] and [19]. To describe these results let $\lambda_1 = 0 < \lambda_2 \leq \lambda_3 \leq \cdots$ denote the eigenvalues of the Laplacian on Ω with Neumann boundary conditions. It was shown in [18] that, for $0 = \lambda_1 < \lambda < \lambda_2$, the set S is a star-shaped smooth manifold of codimension one in the Banach space $E = \{u \in C^{2,\alpha}(\Omega); \frac{\partial u}{\partial n}|_{\partial\Omega} = 0\}$ (see also Church and Timourian [8]) and, for $0 < \lambda < \frac{1}{7}\lambda_2$, the singular set S consists entirely of *fold points* and *cusp points*, the first two singularities in the classification of Whitney and Thom (see also Church, Dancer, and Timourian [9] for the Dirichlet case). Based on this it was shown that, for $0 < \lambda < \frac{1}{12}\lambda_2$, there exists an open set F_3 (containing 0) in the space F such that, if the forcing term f lies in F_3, then equation (2) has exactly three solutions, while if $f \in F_1 := F \backslash \overline{F}_3$ then (2) has exactly one solution.

In [19] it was shown (for certain domains) that there exists some number $\lambda^* \in (\lambda_1, \lambda_2)$ such that there occurs a *forced secondary bifurcation* for $\lambda^* < \lambda < \lambda_2$ in the sense that there are forcing terms for which there exist at least five solutions. Here we will show that in certain cases these singularities can be identified as swallowtail and butterfly singularities.

The paper is organized as follows: In §1 we give the infinite-dimensional characterization of the fold, cusp, swallowtail, and butterfly singularities. For the characterizations of the fold and cusp, see also [1]–[4], [6], [13]. Here we follow the approach of Cafagna and Donati in [6].

In §2 we prove the appearence of swallowtail and butterfly singularities for equation (2) on the interval and the rectangle. Since some of the calculations rely on [19], we recommend that the reader have a copy of this paper at hand.

2. Classification of singularities. In this section we classify the four singularities that will occur later on, namely, fold, cusp, swallowtail, and butterfly. They are the first four singularities that appear in the (finite-dimensional) classification of Thom [20], and also the first four of the so-called Morin singularities, which were classified in the finite-dimensional context by Morin [17]. For illustrations of these singularities, cf. [19].

We make the following general assumptions: Let E and F be Banach spaces such that the inclusions $E \subset F \subset H$ are dense, where H is a suitable Hilbert space. Assume that the nonlinear mapping $\Phi : E \to F$ is smooth and such that its Fréchet derivative $\Phi'(u) : E \to F$ is a Fredholm operator of index zero, with $\dim \mathrm{Ker}\Phi'(u) \leq 1, \forall u \in \mathrm{E}$. Furthermore, assume that $\Phi'(u)$ is symmetric with respect to the inner product in H, that is, denoting this inner product by $(\cdot,\cdot)$, we have $(\Phi'(u)x, y) = (x, \Phi'(u)y), \forall x, y \in \mathrm{E}$.

For nonsymmetric operators the results hold as well; however, some changes will occur in the formulas.

In [6] Cafagna and Donati have given a characterization of the fold and cusp in Banach space, based on the infinite-dimensional version of the Malgrange preparation theorem. Here we use the same approach to characterize the swallowtail and butterfly singularities as well.

Let $u \in S$ be a singular point of Φ. Since $\mathrm{Ker}\Phi'(u) \leq 1$ by assumption, this is equivalent to saying that zero is a simple eigenvalue of $\Phi'(u)$, say $0 = \mu(u)$, with corresponding (normalized) eigenfunction $v \in E$; note that $\mu = \mu(u)$ and $v = v(u)$ depend smoothly on $u \in E$ (cf. [10]). We will see that the singular points can be classified completely in terms of (suitable) derivatives of the function $\mu(u)$. We start with the following local representation lemma (for the general situation, see [6]).

LEMMA 1. *Let $\overline{u} \in E$ be a singular point of Φ as previously specified. Then there exist neighbourhoods $U(\overline{u}) \subset E$, $V(\Phi(\overline{u})) \subset F$, a Banach space X, and smooth diffeomorphisms $\alpha : U \to \alpha(U) \subset R \times X$, $\beta : V \to \beta(V) \subset R \times X$ such that*

$$\Psi = \beta \circ \Phi \circ \alpha^{-1} : \alpha(U) \subset R \times X \to \beta(V) \subset R \times X$$

has the form

$$\Psi(t, x) = (f(t, x); x) \tag{3}$$

Proof. Denote by $v = v(\overline{u})$ the first eigenfunction of $\Phi'(\overline{u})$ with $\|v\| = 1$. Let $F = [v] \oplus F_1$, where F_1 is the orthogonal complement with respect to the inner product in H, and let $P_0 : F \to [v]$, $P_1 : F \to F_1$ denote the respective projections. Let $X = F_1$ and write $u = tv + y \in [v] \oplus P_1 E$. We perform a Lyapunov–Schmidt reduction: For $\overline{t}$ fixed, the operator $P_1\Phi(\overline{t}v + \cdot) : P_1 E \to X$ has a nonvanishing derivative, and hence is locally invertible. Let $\overline{x} = P_1\Phi(\overline{t}v + \overline{y}) = P_1\Phi(\overline{u})$, and denote by $y(t, x)$ the unique solution of

$$P_1\Phi(tv + y) = x \in U(\overline{x}) \subset X,$$

given by the implicit function theorem; $y(t, x)$ depends smoothly on x and t (for $(t; x)$ near $(\overline{t}; \overline{x})$). Now set

$$\begin{aligned} \alpha^{-1} &: U(\overline{t}, \overline{x}) \subset R \times X \to [v] \oplus P_1 E, \\ &(t; x) \to tv + y(t, x). \end{aligned}$$

This mapping is clearly a local diffeomorphism and

$$\Phi \circ \alpha^{-1}(t, x) = (P_0\Phi(tv + y(t, x)); x).$$

Finally, let $\beta : [v] \oplus X \to R \times X$ denote the natural identification $tv + x \to (t; x)$. Then

$$\Psi = \beta \circ \Phi \circ \alpha^{-1} : R \times X \to R \times X$$

has the desired form with

$$f(t, x) = (\Phi(tv + y(t, x)), v). \qquad \square$$

The basic tool for the classification of the singularities is an infinite-dimensional version of the *Malgrange preparation theorem.*

THEOREM 2 (see [16] and [6]). *Let X be a Banach space, U be an open neighbourhood of the origin $(0, \mathbf{0})$ of $R \times X$, and $f : U \to R$ be a smooth function such that*

$$f(0, \mathbf{0}) = \frac{\partial}{\partial t} f(0, \mathbf{0}) = \cdots = \frac{\partial^{k-1}}{\partial t^{k-1}} f(0, \mathbf{0}) = 0, \quad \frac{\partial^k}{\partial t^k} f(0, \mathbf{0}) \neq 0.$$

Then there exist smooth functions $a_j : U \to R$, $j = 0, \dots, k-1$, *vanishing at* $(0, \mathbf{0})$ *such that*

$$t^k = \sum_{j=0}^{k-1} a_j(f(t,z),z) \cdot t^j.$$

Proof. See [16], where the *division theorem* is proved for Banach spaces, and [11] for the proof of the preparation theorem from the division theorem (this proof carries over to the infinite-dimensional situation). We refer also to [6]. □

We now begin with the classification.

DEFINITION 3. *A point* $u \in S$ *is* 1-transverse *if there exists a* $w \in E$ *such that* $\mu'(u)[w] \neq 0$. *We set* $S_1 = \{u \in S; u \text{ is } 1\text{-transverse}\}$.

Clearly, by the implicit function theorem this definition implies that if $u \in S_1$, then S is locally near u a smooth codimension 1 manifold in E.

DEFINITION 4. *Let* $u \in S_1$. *Then* u *is a fold point if*

$$\mu'(u)[v(u)] \neq 0. \tag{4}$$

This is equivalent to saying that $v(u) \notin T_u S$, where $T_u S$ denotes the tangent space to S at u.

We point out that here $v = v(u)$ denotes the eigenfunction of $\Phi'(u)$ corresponding to the eigenvalue $\mu = \mu(u)$. Later, the dependence of $v(u)$ and $\mu(u)$ on the point $u \in E$ will be important and derivatives with respect to u will be taken. To simplify the notation the argument u will sometimes be suppressed.

PROPOSITION 5. Normal form for folds. *If* $\overline{u}$ *is a fold point for* Φ *then there exist neighbourhoods* $U(\overline{u}) \subset E$, $V(\Phi(\overline{u})) \subset F$, *a Banach space* X, *and diffeomorphisms* $\alpha : U \to \alpha(U) \subset R \times X$, $\beta : V \to \beta(V) \subset R \times X$ *such that the following diagram commutes:*

$$\begin{array}{ccc} U \subset E & \xrightarrow{\Phi} & V \subset F \\ \alpha \downarrow & & \downarrow \beta \\ \alpha(U) \subset R \times X & \xrightarrow{\varphi} & \beta(V) \subset R \times X \end{array},$$

where $\varphi : R \times X \to R \times X$, $(t;x) \to (t^2;x)$.

Proof. By Lemma 1 we may consider $\Psi(t,x) = (f(t,x);x)$ with $f(t,x) = (\Phi(tv + y(t,x)), v)$. Note that the singular set of Ψ is given by

$$\tilde{S} = \{(t;x) \in R \times X; f_t(t,x) = 0\},$$

where $f_t := \frac{\partial}{\partial t} f$. Let $(\bar{t};\bar{x})$ be such that $\overline{u} = \bar{t}v + y(\bar{t},\bar{x})$. We then have

$$f_x(\bar{t},\bar{x}) = (\Phi'(\overline{u})y_x(\bar{t},\bar{x}), v) = 0,$$

where $f_x = f'(\bar{t},\bar{x})[(0;x)]$ and $y_x = y'(\bar{t},\bar{x})[(0;x)]$, and

$$f_t(\bar{t},\bar{x}) = (\Phi'(\overline{u})(v + y_t), v) = \mu(\overline{u}) = 0.$$

Note that $y_t := y'(\bar{t},\bar{x})[(1;0)] = 0$ since, by definition (see Lemma 1), $0 = P_1\Phi'(\bar{t}v + y(\bar{t},\bar{x}))(v + y_t) = P_1\Phi'(\bar{t}v + y)y_t$. For f_{tt} we obtain

$$f_{tt}(\bar{t},\bar{x}) = (\Phi''(\overline{u})v^2 + \Phi'(\overline{u})y_{tt}, v) = (\Phi''(\overline{u})v^2, v).$$

On the other hand,

$$\mu_v(\overline{u}) = 2(\Phi'(\overline{u})v, v_v) + (\Phi''(\overline{u})v^2, v) = (\Phi''(\overline{u})v^2, v),$$

and hence

$$f_{tt}(\overline{u}) = \mu_v(\overline{u}) \neq 0 \tag{5}$$

by assumption (here $\mu_v(\overline{u}) := \mu'(\overline{u})[v]$) . Finally, by translation we may assume that $(\overline{t}; \overline{x}) = (0; \mathbf{0})$ and $f(\overline{t}, \overline{x}) = 0$.

Now, applying the preparation Theorem 2 we can write

$$t^2 = a_1(f(t,z), z)t + a_0(f(t,z), z) \tag{6}$$

with $a_0(0, \mathbf{0}) = a_1(0, \mathbf{0}) = 0$. Clearly $\frac{\partial}{\partial t} a_1(f(0, \mathbf{0}), \mathbf{0}) = 0$, and hence the coordinate change $(t; z) \to (t - a_1/2; z)$ is a local diffeomorphism. Finally, setting $b = a_0 + a_1^2/4$, equation (6) becomes

$$t^2 = b(f(t,z), z). \tag{7}$$

Differentiating (7) twice with respect to t we see that $\frac{\partial}{\partial s} b(s,z) \mid_{(s,z)=(0,\mathbf{0})} \neq 0$. Hence the coordinate change in the range $(s; z) \to (b(s,z); z)$ is also a local diffeomorphism and we obtain, in these coordinates,

$$\varphi(t,z) = (t^2; z). \qquad \square$$

REMARK 6. For the particular mapping Φ related to equation (2), conditon (4) is equivalent to

$$F_\lambda(u) := 6 \int_\Omega u v^3(u) d\omega = \mu_v(u) \neq 0. \tag{8}$$

This follows from

$$\begin{aligned} \mu'(u)[v(u)] &= \left(\int_\Omega |\nabla v(u)|^2 - \lambda v(u)^2 + 3u^2 v^2(u) d\omega \right)' [v(u)] \\ &= 2 \int_\Omega (-\Delta v(u) - \lambda v(u) + 3u^2 v(u)) \cdot v'(u)[v(u)] d\omega + 6 \int_\Omega u v^3(u) d\omega \\ &= 6 \int_\Omega u v^3(u) d\omega. \end{aligned}$$

DEFINITON 7. *Suppose that $u \in S_1$ is a 1-transverse singularity which is not a fold, i.e., such that*

$$\mu'(u)[v(u)] = 0. \tag{9}$$

We say that u is 1-1-transverse if there exists a $w \in T_u S_1$ such that

$$(\mu'(u)[v(u)])'[w] \neq 0. \tag{10}$$

Let $S_{1,1} = \{u \in S_1; u \text{ is 1-1-transverse}\}$. By the implicit function theorem it follows that $S_{1,1}$ is locally a smooth submanifold of S_1 of codimension 1.

In what follows we use the notation

$$\mu_{vw}(u) := (\mu'(u)[v(u)])'[w], \quad \mu_{vvw}(u) := ((\mu'(u)[v(u)])'[v(u)])'[w], \text{etc.};$$

note that (albeit the somewhat abusive notation) the function $v = v(u)$ depends on u and has to be differentiated also.

DEFINITION 8. *A point $u \in S_{1,1}$ is a cusp point if*

$$\mu_{vv}(u) := (\mu'(u)[v(u)])'[v(u)] \neq 0. \tag{11}$$

REMARK 9. Geometrically, condition (9) says that $v(u)$ is tangent to S_1, while (11) says that $v(u)$ is not in the tangent space of $S_{1,1}$ (considered as a submanifold of S_1).

PROPOSITION 10 (normal form for cusps). *If $u \in S$ is a cusp point for Φ then there exist neighbourhoods $U(u) \subset E, V(\Phi(u)) \subset F$, a Banach space X, and diffeomorphisms $\alpha : U \to \alpha(U) \subseteq R^2 \times X, \beta : V \to \beta(V) \subseteq R^2 \times X$ such that the following diagram commutes:*

$$\begin{array}{ccc} U(u) \subset E & \xrightarrow{\Phi} & V(\Phi(u)) \subset F \\ \alpha \downarrow & & \downarrow \beta \\ \alpha(U) \subset R^2 \times X & \xrightarrow{\kappa} & \beta(V) \subset R^2 \times X \end{array},$$

where $\kappa(t, s, x) := (t^3 + st; s; x)$.

Proof. By Lemma 1 and the proof of Proposition 5 we may consider $\Psi(t, x) = (f(t, x); x)$ with

$$f_t(\bar{t}, \bar{x}) = f_x(\bar{t}, \bar{x}) = 0,$$

and by (5) and assumption we see that $f_{tt}(\bar{t}, \bar{x}) = 0$ also. We show that $f_{ttt}(\bar{t}, \bar{x}) \neq 0$. In fact

$$f_{ttt}(\bar{t}, \bar{x}) = (\Phi^{(3)}(\bar{u})v^3 + 3\Phi''(\bar{u})vy_{tt} + \Phi'(\bar{u})y_{ttt}, v).$$

On the other hand, using the fact that $\Phi'(\bar{u})$ is symmetric and hence $(\Phi''(\bar{u})v^2, v_v) = (\Phi''(\bar{u})vv_v, v)$, we have

$$\begin{aligned} \mu_{vv}(\bar{u}) = (\Phi^{(3)}(\bar{u})v^3, v) + 3(\Phi''(\bar{u})vv_v, v) + 2(\Phi''(\bar{u})v^2, v_v) \\ + 2(\Phi'(\bar{u})v_v, v_v) + 2(\Phi'(\bar{u})v, v_{vv}). \end{aligned}$$

Here $v_v := v'(\bar{u})[v]$ and $v_{vv} = (v'(\bar{u})[v(\bar{u})])'[v]$. Since $\mu(\bar{u}) = \mu_v(\bar{u}) = 0$ by assumption, we have

$$\Phi''(\bar{u})v^2 + \Phi'(\bar{u})v_v = \mu(\bar{u})v_v + \mu_v(\bar{u})v = 0. \tag{12}$$

Taking the t-derivative of $P_1\Phi'(tv + y(t, \bar{x}))(v + y_t) = 0$ in $t = \bar{t}$ and recalling that $y_t = 0$ in $t = \bar{t}$, we get

$$P_1\Phi''(\bar{t}v + y(\bar{t}, \bar{x}))v^2 + P_1\Phi'(\bar{t}v + y(\bar{t}, \bar{x}))y_{tt} = 0,$$

and since $\Phi''(\bar{u})v^2 \perp v$, $P_1 : F \to [v]^\perp$ and $\bar{u} = \bar{t}v + y(\bar{t}, \bar{x})$, we conclude that $y_{tt} = v_v$. Hence

$$f_{ttt}(\bar{t}, \bar{x}) = \mu_{vv}(\bar{u}) \neq 0. \tag{13}$$

We also need to verify the 1-transversality condition for Ψ, that is, $f_{tx}(\bar{t},\bar{x}) \not\equiv 0$. In fact we have

$$f_{tx}(\bar{t},\bar{x}) = (\Phi''(\bar{u})vy_x + \Phi'(\bar{u})y_{tx}, v).$$

By (10) there exists $z \in P_1E$ such that

$$0 \neq \mu_{vz}(\bar{u}) = (\Phi''(\bar{u})vz, v) + 2(\Phi'(\bar{u})v_z, v).$$

Set $x = P_1\Phi'(\bar{u})z$; then $y_x(\bar{t},\bar{x}) = z$, since $P_1\Phi'(\bar{u})y_w = w$ for all $w \in P_1E$. Hence we conclude

$$f_{tx}(\bar{t},\bar{x}) = \mu_z(\bar{u}) \neq 0. \tag{14}$$

Again, by translation we assume that $(\bar{t};\bar{x}) = (0;\mathbf{0})$ and $f(\bar{t},\bar{x}) = 0$.

By the preparation theorem we find smooth functions $a_j : R\times X \to R$, $j = 0,1,2$, vanishing in $(0;\mathbf{0})$ and such that

$$t^3 = a_2(f(t,z),z)t^2 + a_1(f(t,z),z)t + a_0(f(t,z),z). \tag{15}$$

Since $\frac{\partial}{\partial t}a_2(f(0,\mathbf{0}),\mathbf{0}) = 0$, the coordinate change $(t;z) \to (t - \frac{1}{3}a_2;z)$ is a local diffeomorphism, and setting $b_1 = -\frac{1}{3}\cdot a_2^2 - a_1$ and $b_0 = \frac{2}{27}\cdot a_2^3 + \frac{1}{3}\cdot a_1a_2 + a_0$, equation (15) is transformed into

$$t^3 + b_1(f(t,z),z)t = b_0(f(t,z),z). \tag{16}$$

Let L be a one-dimensional subspace of X and let $X = L \oplus Y$. Furthermore, let $P : X \to Y$ denote the canonical projection onto Y. Define the local coordinate changes

$$\begin{array}{ccc} R\times X & \xrightarrow{\Psi} & R\times X \\ q_1\downarrow & & \downarrow q_2 \\ R\times L\times Y & \xrightarrow{\kappa} & R\times L\times Y \end{array},$$

where

$$q_1(t,x) = (t;\tilde{b}_1(t,x);Px), \quad \text{with } \tilde{b}_1(t,x) = b_1(f(t,x),x),$$
$$q_2(s,x) = (b_0(s,x);b_1(s,x);Px).$$

Clearly, $\kappa = q_2 \circ \Psi \circ q_1^{-1}$ then has the form

$$\kappa(t,w,x_1) = (t^3 + tw; w; x_1). \tag{17}$$

It remains to verify that q_1 and q_2 are local diffeomorphisms. For this it suffices to show that

$$\frac{\partial}{\partial s}b_0(0,\mathbf{0}) \neq 0\ , \quad \frac{\partial}{\partial x}b_1(0,\mathbf{0}) \not\equiv 0.$$

Calculating the third derivative of (16) with respect to t in $(0,\mathbf{0})$ yields $6 = \frac{\partial}{\partial s}b_0(0,\mathbf{0})\cdot(\partial^3/\partial t^3)f(0,\mathbf{0})$. Calculating the mixed second derivative of (16) with respect to t and x in $(0,\mathbf{0})$ yields $\frac{\partial}{\partial x}b_1(0,\mathbf{0}) = \frac{\partial}{\partial s}b_0(0,\mathbf{0})\cdot(\partial^2/\partial t\partial x)f(0,\mathbf{0})$. The second mixed derivative of f in $(0,\mathbf{0})$ is different from zero by the 1-transversality of Ψ. □

Remark 11. For equation (2) , condition (11) is equivalent to

$$C_\lambda(u) := 6\int_\Omega v^4(u)d\omega + 18\int_\Omega uv^2(u)v_v d\omega = \mu_{vv}(u) \neq 0. \tag{18}$$

This is obtained by differentiating $F_\lambda(u)$ as given by (8) in direction v. The expression $v_v = v'(u)[v(u)]$ can be calculated by taking the derivative of $(-\Delta - \lambda + 3u^2)v(u) = \mu(u)v(u) = 0$ in direction $v = v(u)$. We obtain

$$(-\Delta - \lambda + 3u^2)v_v + 6uv^2 = \mu_v(u)v + \mu(u)v_v = 0 \tag{19}$$

by (9), since $u \in S$. By (8) condition (9) implies that uv^2 is normal to v, and hence we get

$$v_v = (-\Delta - \lambda - 3u^2)^{-1}(-6uv^2). \tag{20}$$

Using the spectral representation of $\Phi'(u)$ and (20), $C_\lambda(u)$ can be alternatively written as

$$C_\lambda(u) = 6\int_\Omega v^4(u)d\omega - 108\sum \frac{1}{\mu_i(u)}\left(\int_\Omega uv^2(u)v_i(u)d\omega\right)^2, \tag{21}$$

where the sum is extended over all eigenvalues $\mu_i(u) \neq \mu(u)$ of $\Phi'(u)$ ($v_i(u)$ are the eigenfunctions corresponding to $\mu_i(u)$).

We continue with the classification in the (by now) obvious way.

DEFINITION 12. *Suppose that $u \in S_{1,1}$ is a 1-1-transverse singularity which is not a cusp, i.e., such that*

$$\mu_{vv}(u) := (\mu'(u)[v(u)])'[v(u)] = 0. \tag{22}$$

We say that u is 1-1-1-transverse if there exists a $w \in T_uS_{1,1}$ such that

$$\mu_{vvw}(u) := ((\mu'(u)[v(u)])'[v(u)])'[w] \neq 0. \tag{23}$$

Let $S_{1,1,1} = \{u \in S_{1,1};\ u \text{ is 1-1-1-transverse}\}$.

By the implicit function theorem it follows that $S_{1,1,1}$ is a locally smooth submanifold of codimension 1 of $S_{1,1}$.

DEFINITION 13. *A point $u \in S_{1,1,1}$ is a swallowtail singularity if*

$$\mu_{vvv}(u) := ((\mu'(u)[v(u)])'[v(u)])'[v(u)] \neq 0. \tag{24}$$

PROPOSITION 14 (normal form for swallowtails). *If $u \in S$ is a swallowtail singularity for Φ then there exist neighbourhoods $U(u) \subset E, V(\Phi(u)) \subset F$, a Banach space X, and local diffeomorphisms $\alpha : U \to R^3 \times X$, $\beta : V \to R^3 \times X$ such that*

$$\sigma := \beta \circ \Phi \circ \alpha^{-1} : \alpha(U) \subset R^3 \times X \to \beta(V) \subset R^3 \times X$$

has the form

$$\sigma(r, s, t, x) = (t^4 + st^2 + rt; s; r; x).$$

Proof. We can consider $\Psi(t, x) = (f(t, x); x)$ with

$$f_t(\bar{t}, \bar{x}) = f_{tt}(\bar{t}, \bar{x}) = f_x(\bar{t}, \bar{x}) = 0,$$

and by (13) and assumption $f_{ttt}(\bar{t}, \bar{x}) = 0$ also. We show that $f_{tttt}(\bar{t}, \bar{x}) \neq 0$. In fact, from the expression for f_{ttt} we deduce

$$\begin{aligned} f_{tttt}(\bar{t}, \bar{x}) = (\Phi^{(4)}(\bar{u})v^4 + 6\Phi^{(3)}(\bar{u})v^2 y_{tt} + 3\Phi''(\bar{u})y_{tt}^2 \\ + 4\Phi''(\bar{u})vy_{ttt} + \Phi'(\bar{u})y_{tttt}, v). \end{aligned}$$

On the other hand,

$$\mu_{vvv}(\overline{u}) = (\Phi^{(4)}(\overline{u})v^4, v) + 9(\Phi^{(3)}(\overline{u})v^3, v_v) + 7(\Phi''(\overline{u})v^2, v_{vv}) + 12(\Phi''(\overline{u})v_v^2, v) + 6(\Phi'(\overline{u})v_v, v_{vv}) + 2(\Phi'(\overline{u})v_{vvv}, v).$$

One verifies as before that $y_{ttt} = v_{vv}$ and $y_{tttt} = v_{vvv}$. Using (12),

$$\Phi''(\overline{u})v^2 + \Phi'(\overline{u})v_v = \mu(\overline{u})v_v + \mu_v(\overline{u})v = 0,$$

and its derivative in direction v,

$$\Phi^{(3)}(\overline{u})v^3 + 3\Phi''(\overline{u})vv_v + \Phi'(\overline{u})v_{vv} = \mu(\overline{u})v_{vv} + 2\mu_v(\overline{u})v_v + \mu_{vv}(\overline{u})v = 0, \tag{25}$$

we conclude that

$$f_{tttt}(\overline{t}, \overline{x}) = \mu_{vvv}(\overline{u}) \neq 0. \tag{26}$$

The 1-transversality holds as in Proposition 10. For the 1-1-transversality of Ψ we have to show that $f_{ttx}(\overline{t}, \overline{x}) \neq 0$ for some $x \in T_{(\overline{t},\overline{x})}\tilde{S}$ ($\Longleftrightarrow f_{tx}(\overline{t}, \overline{x}) = 0$). We have

$$f_{ttx}(\overline{t}, \overline{x}) = (\Phi^{(3)}(\overline{u})v^2 y_x + \Phi''(\overline{u})y_x y_{tt} + 2\Phi''(\overline{u})vy_{tx}, v)$$

and

$$\mu_{vw}(\overline{u}) = (\Phi^{(3)}(\overline{u})v^2 w, v) + 3(\Phi''(\overline{u})v^2, v_w) + 2(\Phi''(\overline{u})wv_v, v) + 2(\Phi'(\overline{u})v_{vw}, v) + 2(\Phi'(\overline{u})v_v, v_w).$$

For $w \in T_{\overline{u}}S$ we have $\mu_w(\overline{u}) = \mu(\overline{u}) = 0$, and hence

$$\Phi''(\overline{u})vw + \Phi'(\overline{u})v_w = \mu(\overline{u})v_w + \mu_w(\overline{u})v = 0. \tag{27}$$

Choosing $\overline{x}$ such that $y_x(\overline{t}, \overline{x}) = w$, i.e. $\overline{x} = \Phi'(\overline{u})w$, we find $y_{tx}(\overline{t}, \overline{x}) = v_w(\overline{u})$ by equations (27) and $P_1\Phi''(\overline{u})vy_x + \Phi'(\overline{u})y_{tx} = 0$. This, together with (12) and (27), implies

$$f_{ttx}(\overline{t}, \overline{x}) = \mu_{vw}(\overline{u}) \tag{28}$$

with $f_{tx}(\overline{t}, \overline{x}) = 0 \Longleftrightarrow \mu_w(\overline{u}) = 0$ by (14) .

By the preparation theorem we can now write

$$t^4 = a_3(f(t,x), x) \cdot t^3 + a_2(f(t,x), x) \cdot t^2 + a_1(f(t,x), x) \cdot t + a_0(f(t,x), x), \tag{29}$$

where $a_i : R \times X \to R, i = 0, 1, 2, 3$, are smooth functions vanishing in (0,**0**). The coordinate change $(t; x) \to (t - a_3/4; x)$ is a local diffeomorphism, since $\frac{\partial}{\partial t}a_3(f(0, \mathbf{0}), \mathbf{0}) = 0$. Now, setting

$$b_2 = -\frac{3}{8}a_3^2 - a_2, \quad b_1 = -\frac{1}{8}a_3^3 - \frac{1}{2}a_2a_3 - a_1,$$
$$b_0 = \frac{3}{4^4}a_3^4 + \frac{1}{16}a_2a_3^2 + \frac{1}{4}a_1a_3 + a_0,$$

equation (29) is transformed into

$$t^4 + b_2(f(t,x),x)\cdot t^2 + b_1(f(t,x),x)\cdot t = b_0(f(t,x),x). \tag{30}$$

Let L_1 and L_2 be linearly independent, one-dimensional subspaces of X, and let $X = L_1 \oplus L_2 \oplus Y$; furthermore, let $P : X \to Y$ denote the canonical projection. Define the local coordinate changes

$$\begin{array}{ccc} R\times X & \xrightarrow{\Psi} & R\times X \\ q_1 \downarrow & & \downarrow q_2 \\ R\times L_2\times L_1\times Y & \xrightarrow{\tilde{\Psi}} & R\times L_2\times L_1\times Y \end{array},$$

where

$$q_1(t,x) = (t;\tilde{b}_2(t,x);\tilde{b}_1(t,x);Px) \quad \text{with } \tilde{b}_i(t,x) := b_i(f(t,x),x),\ i=1,2,$$
$$q_2(s,x) = (b_0(s,x);b_2(s,x);b_1(s,x);Px).$$

Then $\sigma = \tilde{\Psi}$ has the form

$$\sigma(t,s,r,y) = (t^4 + s\cdot t^2 + r\cdot t; s; r; y).$$

It remains to show that q_1 and q_2 are local diffeomorphisms. For this it suffices to show that

$$\frac{\partial}{\partial s}b_0(0,\mathbf{0}) \neq 0, \quad \frac{\partial}{\partial x}b_1(0,\mathbf{0}) \not\equiv 0, \quad \frac{\partial}{\partial x}b_2(0,\mathbf{0}) \not\equiv 0.$$

Taking the fourth derivative of (30) with respect to t in (0,**0**) yields $24 = \frac{\partial}{\partial s}b_0(0,\mathbf{0})\cdot(\partial^4/\partial t^4)f(0,\mathbf{0})$, and hence $\frac{\partial}{\partial s}b_0(0,\mathbf{0}) \neq 0$ by assumption. Taking the mixed second-order derivative with respect to t and x of (30) in (0,**0**) gives $\frac{\partial}{\partial x}b_1(0,\mathbf{0}) = \frac{\partial}{\partial s}b_0(0,\mathbf{0})\cdot(\partial^2/\partial t\partial x)f(0,\mathbf{0})$, and since $(\partial^2/\partial t\partial x)f(0,\mathbf{0}) \not\equiv 0$ by the 1-transversality assumption, we conclude that $\frac{\partial}{\partial x}b_1(0,\mathbf{0}) \not\equiv 0$. Finally, taking the mixed third-order derivative $\partial^3/\partial t^2\partial x$ of (30) in (0,**0**) yields $2b_{2,x}(0,\mathbf{0}) + 2b_{1,s}(0,\mathbf{0})\cdot f_{tx}(0,\mathbf{0}) = b_{0,s}(0,\mathbf{0})\cdot f_{ttx}(0,\mathbf{0})$. By the 1-1-transversality we find an x such that $f_{tx} = 0$ and $f_{ttx} \neq 0$; hence $b_{2,x}(0,\mathbf{0}) \not\equiv 0$. □

Remark 15. For equation (2) condition (24) is equivalent to

$$Sw_\lambda(u) := 60\cdot\int_\Omega v^3 v_v + 90\cdot\int_\Omega u\cdot v v_v^2 = \mu_{vvv}(u) \neq 0, \tag{31}$$

where v_v is given by (20). In fact, this follows from calculating the derivative of $C_\lambda(u) = \mu_{vv}(u)$ (given by (18)) in direction v;

$$C'_\lambda(u)[v] = \mu_{vvv}(u) = 42\cdot\int_\Omega v^3 v_v + 36\cdot\int_\Omega u v v_v^2 + 18\cdot\int_\Omega u v^2 v_{vv}.$$

The term v_{vv} is obtained by differentiating (19) in direction v and noting that $\mu_v(u) = \mu_{vv}(u) = 0$ by assumption:

$$(-\Delta - \lambda + 3u^2)v_{vv} = -6(v^3 + 3uvv_v). \tag{32}$$

With (32), and using (20), the last term in the previous expression can be written as $18\int_0^1 uv^2v_{vv} = 18\int_0^1 v_v(v^3+3uvv_v)$, hence (31) holds.

DEFINITION 16. *Suppose that $u \in S_{1,1,1}$ is a* 1-1-1-*transverse singularity which is not a swallowtail, i.e., assume that*

$$\mu_{vvv}(u) = ((\mu'(u)[v(u)])'[v(u)])'[v(u)] = 0. \tag{33}$$

We say that u is 1-1-1-1-*transverse if there exists a $w \in T_uS_{1,1,1}$ such that*

$$\mu_{vvvw}(u) = (((\mu'(u)[v(u)])'[v(u)])'[v(u)])'[w] \neq 0. \tag{34}$$

Let $S_{1,1,1,1} = \{u \in S_{1,1,1};\ u \text{ is 1-1-1-1-transverse}\}$. The implicit function theorem implies that $S_{1,1,1,1}$ is a smooth submanifold of codimension 1 of $S_{1,1,1}$.

DEFINITION 17. *A point $u \in S_{1,1,1,1}$ is a butterfly singularity if*

$$\mu_{vvvv}(u) = (((\mu'(u)[v(u)])'[v(u)])'[v(u)])'[v(u)] \neq 0. \tag{35}$$

PROPOSITION 18 (normal form for butterflies). *If $\overline{u} \in S$ is a butterfly singularity then there exist neighbourhoods $U(\overline{u}) \subset E$, $V(\Phi(\overline{u})) \subset F$, a Banach space X, and local diffeomorphisms $\alpha : U \to \alpha(U) \subseteq R^4 \times X$, $\gamma : V \to \gamma(V) \subseteq R^4 \times X$ such that*

$$\beta := \gamma \circ \Phi \circ \alpha^{-1} : \alpha(U) \subseteq R^4 \times X \to \gamma(V) \subseteq R^4 \times X$$

has the form

$$\beta(r,s,t,q,x) = (t^5 + st^3 + rt^2 + qt; s; r; q; x).$$

Proof. We can consider $\Psi(t,x) = (f(t,x);x)$ with

$$f_t(\overline{t},\overline{x}) = f_{tt}(\overline{t},\overline{x}) = f_{ttt}(\overline{t},\overline{x}) = 0,$$

and by (26) and assumption,

$$f_{tttt}(\overline{t},\overline{x}) = \mu_{vvv}(\overline{u}) = 0.$$

We show that

$$f_{ttttt}(\overline{t},\overline{x}) = \mu_{vvvv}(\overline{u}) \neq 0. \tag{36}$$

In fact

$$\begin{aligned} f_{ttttt}(\overline{t},\overline{x}) = (&\Phi^{(5)}(\overline{u})v^5 + 10\Phi^{(4)}(\overline{u})v^3y_{tt} + 15\Phi^{(3)}(\overline{u})vy_{tt}^2 \\ &+ 10\Phi^{(3)}(\overline{u})v^2y_{ttt} + 10\Phi''(\overline{u})y_{tt}y_{ttt} \\ &+ 5\Phi''(\overline{u})vy_{tttt} + \Phi'(\overline{u})y_{ttttt}, v). \end{aligned}$$

For $\mu_{vvvv}(\overline{u})$ we calculate

$$\begin{aligned} \mu_{vvvv}(\overline{u}) =&(\Phi^{(5)}(\overline{u})v^5, v) + 14(\Phi^{(4)}(\overline{u})v^4, v_v) + 39(\Phi^{(3)}(\overline{u})v^2, v_v^2) \\ &+ 16(\Phi^{(3)}(\overline{u})v^3, v_{vv}) + 44(\Phi''(\overline{u})vv_v, v_{vv}) + 9(\Phi''(\overline{u})v^2, v_{vvv}) \\ &+ 12(\Phi''(\overline{u})v_v^2, v_v) + 6(\Phi'(\overline{u})v_{vv}, v_{vv}) + 8(\Phi'(\overline{u})v_v, v_{vvv}) + 2(\Phi'(\overline{u})v_{vvvv}, v). \end{aligned}$$

Taking the v-derivative of (25) we have

$$\begin{aligned} &\Phi^{(4)}(\overline{u})v^4 + 6\Phi^{(3)}(\overline{u})v^2v_v + 4\Phi''(\overline{u})vv_{vv} + 3\Phi''(\overline{u})v_v^2 + \Phi'(\overline{u})v_{vvv} \\ &= \mu(\overline{u})v_{vvv} + 3\mu_v(\overline{u})v_{vv} + 3\mu_{vv}(\overline{u})v_v + \mu_{vvv}(\overline{u})v = 0. \end{aligned} \tag{37}$$

Using (37), (25), (12), and the assumption that $\overline{u}$ is a butterfly singlarity, we infer (36).

Observe that $x \in T_{(\overline{t},\overline{x})}\tilde{S}_1$ and $x \in T_{(\overline{t},\overline{x})}\tilde{S}_{1,1}$, since

$$\mu_w(\overline{u}) = 0 \Longleftrightarrow f_{tx}(\overline{t},\overline{x}) = 0,$$
$$\mu_{vw}(\overline{u}) = 0 \Longleftrightarrow f_{ttx}(\overline{t},\overline{x}) = 0.$$

Finally, we can asssume that $(\overline{t};\overline{x}) = (0;\mathbf{0})$ and $f(\overline{t},\overline{x}) = 0$.

By the preparation theorem we can now write

$$\begin{aligned} t^5 = {} & a_4(f(t,x),x)\cdot t^4 + a_3(f(t,x),x)\cdot t^3 \\ & + a_2(f(t,x),x)\cdot t^2 + a_1(f(t,x),x)\cdot t + a_0(f(t,x),x), \end{aligned} \tag{38}$$

where $a_i : R\times X \to R, i = 0,1,\ldots,4$, are smooth functions vanishing in $(0,\mathbf{0})$. The coordinate change $(t;x) \to (t - a_4/5;x)$ is a local diffeomorphism, since $\frac{\partial}{\partial t}a_4(f(0,\mathbf{0}),\mathbf{0}) = 0$. Setting

$$b_3 = -\frac{2}{5}a_4^2 - a_3, \quad b_2 = -\frac{4}{5^2}a_4^3 - \frac{3}{5}a_3a_4 - a_2,$$
$$b_1 = -\frac{3}{5^3}a_4^4 - \frac{3}{5^2}a_3a_4^2 - \frac{2}{5}a_2a_4 - a_1,$$
$$b_0 = \frac{4}{5^5}a_4^5 + \frac{1}{5^3}a_3a_4^3 + \frac{1}{5^2}a_2a_4^2 + \frac{1}{5}a_1a_4 + a_0,$$

equation (38) is transformed into

$$t^5 + b_3(f(t,x),x)\cdot t^3 + b_2(f(t,x),x)\cdot t^2 + b_1(f(t,x),x)\cdot t = b_0(f(t,x),x). \tag{39}$$

Let L_1, L_2, and L_3 be linearly independent, one-dimensional subspaces of X, and $X = L_1 \oplus L_2 \oplus L_3 \oplus Y$, and denote the canonical projection from X onto Y by $P : X \to Y$. Define the local coordinate changes

$$\begin{array}{ccc} U \subset R\times X & \xrightarrow{\Phi} & V \subset R\times X \\ q_1 \downarrow & & \downarrow q_2 \\ q_1(U) \subset R\times L_3\times L_2\times L_1\times Y & \xrightarrow{\tilde{\Phi}} & q_2(V) \subset R\times L_3\times L_2\times L_1\times Y \end{array},$$

where

$$q_1(t,x) = (t;\tilde{b}_3(t,x);\tilde{b}_2(t,x);\tilde{b}_1(t,x);Px), \quad \text{with } \tilde{b}_i(t,x) := b_i(f(t,x),x),\ i = 1,2,3,$$
$$q_2(s,x) = (b_0(s,x);b_3(s,x);b_2(s,x);b_1(s,x);Px).$$

Then $\beta = \tilde{\Phi}$ has the form

$$\beta(t,s,r,q,y) = (t^5 + s\cdot t^3 + r\cdot t^2 + q\cdot t;s;r;q;y).$$

To establish that q_1 and q_2 are local diffeomorphisms, we show that

$$\frac{\partial}{\partial s}b_0(0,\mathbf{0}) \neq 0, \quad \frac{\partial}{\partial x}b_i(0,\mathbf{0}) \not\equiv 0, \quad i = 1,2,3.$$

In fact, taking the fifth derivative of (39) with respect to t in $(0;\mathbf{0})$ yields $120 = \frac{\partial}{\partial s}b_0(0,\mathbf{0})\cdot(\partial^5/\partial t^5)f(0,\mathbf{0})$, and hence $\frac{\partial}{\partial s}b_0(0,\mathbf{0}) \neq 0$ by hypothesis. The second and

third inequalities then follow as they do for the swallowtail by taking the mixed second and third derivatives of (39) with respect to t and x in $(0,\mathbf{0})$; the fourth inequality follows by taking the mixed fourth derivative $\partial^4/\partial x\partial t^3$ of (39):

$$6b_{3,x}(0,\mathbf{0}) + 6b_{2,s}(0,\mathbf{0})f_{tx}(0,\mathbf{0}) + 3b_{1,s}(0,\mathbf{0})f_{ttx}(0,\mathbf{0}) = b_{0,s}(0,\mathbf{0})f_{tttx}(0,\mathbf{0}).$$

By the 1-1-1-transversality we find an $x \in X$ with $f_{tx}(0,\mathbf{0}) = f_{ttx}(0,\mathbf{0}) = 0$ and $f_{tttx}(0,\mathbf{0}) \neq 0$; hence $b_{s,x}(0,\mathbf{0}) \neq 0$. □

Remark 19. For equation (2) condition (35) is equivalent to

$$\begin{aligned} B_\lambda(u) &:= 270\int_\Omega v^2v_v^2d\omega + 60\int_\Omega v^3v_{vv}d\omega + 180\int_\Omega uvv_vv_{vv}d\omega + 90\int_\Omega uv_v^3d\omega \\ &= \mu_{vvvv}(u) \neq 0\ . \end{aligned} \tag{40}$$

This follows from differentiating (31) in direction v.

For easier reference, we recall here all the conditions that must hold so that a given point $u \in E$ is a butterfly singularity for the mapping $\Phi : E \to F$, $\Phi(u) = -\Delta u - \lambda u + u^3$.

(b_1) u is singular point (i.e., $u \in S$):
$\mu(u) = \int_0^1 v'^2d\omega - \lambda\int_0^1 v^2d\omega + 3\int_0^1 u^2v^2d\omega = 0$.

(b_2) u is 1-transverse (i.e., $u \in S_1$):
$\exists z \in E$ such that $\mu_z(u) = 6\int_0^1 uv^2zd\omega \neq 0$.

(b_3) u is not a fold singularity:
$\mu_v(u) = 6\int_0^1 uv^3d\omega = 0$.

(b_4) u is 1-1-transverse (i.e., $u \in S_{1,1}$):
$\exists w \in E$ with $\mu_w(u) = 6\int_0^1 uv^2wd\omega = 0$ (i.e., $w \in T_uS_1$) such that
$\mu_{vw}(u) = 6\int_0^1 wv^3d\omega + 18\int_0^1 uv^2v_wd\omega \neq 0$.

(b_5) u is not a cusp singularity:
$\mu_{vv}(u) = 6\int_0^1 v^4d\omega + 18\int_0^1 uv^2v_vd\omega = 0$.

(b_6) u is 1-1-1-transverse (i.e., $u \in S_{1,1,1}$):
$\exists y \in E$ with $\int_0^1 uv^2yd\omega = 0$ and $\int_0^1 yv^3 + 3uv^2v_yd\omega = 0$ (i.e., $y \in T_uS_{1,1}$) such that
$u_{vvy}(u) = 24\int_0^1 v^3v_yd\omega + 36\int_0^1 yv^2v_vd\omega + 72\int_0^1 uvv_yv_vd\omega + 18\int_0^1 uyv_v^2d\omega \neq 0$
(this is obtained by differentiating $\mu_{vv}(u)$ in direction y and using
$v_{vy} = -6\Phi(u)^{-1}(uyv_v + yv^2 + 2uvv_y)$ and $v_v = -6\Phi'(u)^{-1}uv^2$).

(b_7) u is not a swallowtail singularity:
$\mu_{vvv}(u) = 60\int_0^1 v^3v_vd\omega + 90\int_0^1 uvv_v^2d\omega = 0$.

(b_8) $\mu_{vvvv}(u) = 270\int_0^1 v^2v_v^2d\omega + 60\int_0^1 v^3v_{vv}d\omega + 180\int_0^1 uvv_vv_{vv}d\omega + 90\int_0^1 uv_v^3d_\omega \neq 0$; this implies in particular that u is 1-1-1-1-transverse (i.e., $u \in S_{1,1,1,1}$).

3. The elliptic boundary value problem (2). Here we prove that for problem (2) with $\Omega = (0,1)$ and for parameter values $\lambda > \lambda^*$ and near λ^* (where $\lambda^* := \lambda_2 \cdot 2/\pi^2$, cf. [19, Prop. 12]), the singular set S contains butterfly singularities.

THEOREM 20. *Assume that $\Omega = (0,1)$ and let $\Phi : E \to F$ be given by $\Phi(u) = -u'' - \lambda u + u^3$ with $E = \{u \in C^2(0,1), u'(0) = u'(1) = 0\}$ and $F = C(0,1)$. Let $\lambda^* = \lambda_2 2/\pi^2$. Then there exists an $\epsilon > 0$ such that, for $\lambda^* < \lambda < \lambda^* + \epsilon$, the singular set contains a butterfly singularity.*

Proof. As in [19, p. 797], let

$$z(x) = \begin{cases} 1 & \text{if } x \in (0, \frac{1}{2}], \\ -1 & \text{if } x \in (\frac{1}{2}, 1). \end{cases}$$

Note that $z \in L^p(0,1), \forall p \geq 1$ but not in F; however, the operator

$$\Phi'(tz) := -\frac{d^2}{dx^2} - \lambda + 3(tz)^2 = -\frac{d^2}{dx^2} - \lambda + 3t^2 : E \to F$$

with $t \in R^+$ is well defined with corresponding eigenvalues $\mu_i(tz) = \lambda_i - \lambda + 3t^2$ (λ_i, $i \in N$, denotes the Neumann eigenvalues of $-d^2/dx^2$).

The proof of the theorem proceeds by approximation. It follows from the previous section that the property that u is a butterfly singularity of Φ is definable purely in terms of $\Phi'(u)$ and is given by integral conditions. We set $\lambda = \lambda^*$ and $\bar{t} = (\lambda^*/3)^{1/2}$ and verify that $\Phi'(\bar{t}z)$ satisfies these conditions. Then we show that there are $z_\epsilon \in E$ with $z_\epsilon \to z$ in $L^p(0,1)$, $\forall p \geq 1$, and $t_\epsilon \to \bar{t}$ such that the integral conditions are also satisfied in $u_\epsilon = t_\epsilon z_\epsilon$. Thus Φ has a butterfly singularity at $u_\epsilon = t_\epsilon z_\epsilon$.

First note that for $\lambda = \lambda^*$ and $\bar{t} = (\lambda^*/3)^{1/2}$ we have $\Phi'(\bar{t}z)[v(\bar{t}z)] = \mu(\bar{t}z)v(\bar{t}z) = 0$ (hence (b_1)), and since $\mu'(\bar{t}z)[z] = 6\int_0^1 \bar{t}z^2v^2(\bar{t}z)dx > 0$, the 1-transversality condition (b_2) is satisfied in $\bar{t}z$.

Next, note that $\mu'(\bar{t}z)[v(\bar{t}z)] = 6\int_0^1 \bar{t}zv^3(\bar{t}z)dx = 0$, since $v(tz) \equiv 1$; hence the condition b_3 is satisfied.

We claim that in $\bar{t}z$ the 1-1-transversality condition (b_4) is satisfied. In fact, let

$$w = \begin{cases} 1, & x \in (0, \frac{1}{4}] \cup (\frac{3}{4}, 1), \\ -1, & x \in (\frac{1}{4}, \frac{3}{4}]. \end{cases}$$

Then we have $\mu_w(\bar{t}z) = \int_0^1 \bar{t}zv^2(\bar{t}z)wdx = \int_0^1 \bar{t}zwdx = 0$. Furthermore, writing $v = v(\bar{t}z)$ and $v_w = v'(\bar{t}z)[w]$,

$$6\int_0^1 v^3wdx + 18\int_0^1 \bar{t}zv^2v_w = 18\int_0^1 \bar{t}zv_wdx.$$

Note that from $-v_w'' - \lambda v_w + 3(\bar{t}z)^2v_w + 6\bar{t}zwv = \mu(\bar{t}z)v_w + \mu_w(\bar{t}z)v = 0$, we get (using $\lambda = 3(\bar{t}z)^2$ and $v \equiv 1$) $v_w'' = 6\bar{t}zw$ with $v_w'(0) = v_w'(1) = 0$. This yields

$$v_w = 3\bar{t}z \cdot \begin{cases} x^2 - \frac{1}{8}, & x \in (0, \frac{1}{4}), \\ -(\frac{1}{2} - x)^2, & x \in (\frac{1}{4}, \frac{1}{2}), \\ (\frac{1}{2} - x)^2, & x \in (\frac{1}{2}, \frac{3}{4}), \\ -(1-x)^2 + \frac{1}{8}, & x \in (\frac{3}{4}, 1). \end{cases}$$

With this representation for v_w we clearly obtain $\int^1 zv_wdx = -\int_0^1 |v_w|dx \neq 0$, and hence the 1-1-transversality condition (b_4) is verified in $\bar{t}z$.

Next we note that in the point $\bar{t}z$ the cusp condition is not verified; that is, we have (b_5):

$$C_{\lambda^*}(\bar{t}z) = 6\int_0^1 v^4dx + 18\int_0^1 \bar{t}zv^2v_vdx = 0.$$

In fact, from $v_v''(x) = 6\bar{t}z$, $v_v'(0) = v_v'(1) = 0$, we get

$$v_v(x) = \begin{cases} 3\bar{t}(x^2 - \frac{1}{4}), & 0 < x \leq \frac{1}{2}, \\ 3\bar{t}(-(1-x)^2 + \frac{1}{4}), & (\frac{1}{2}) < x < 1, \end{cases} \tag{41}$$

from which the claim follows by calculation (cf. [19, Prop. 12]). We remark that the special choice of λ^* enters here.

We show that the 1-1-1-transversality condition (b_6) holds in $\bar{t}z$. For this we set $y = z \cdot w$. Then $\int_0^1 zv^2 y dx = \bar{t}\int_0^1 w dx = 0$ and $\int_0^1 yv^3 + 3\bar{t}zv^2 v_y dx = 0$, since v_y is symmetric with respect to $x = \frac{1}{2}$. From $v_y'' = 6\bar{t}zvy = 6\bar{t}w$ we infer that

$$v_y = 3\bar{t} \cdot \begin{cases} x^2 - \frac{1}{16}, & x \in (0, \frac{1}{4}), \\ -(x - \frac{1}{2})^2 + \frac{1}{16}, & x \in (\frac{1}{4}, \frac{3}{4}), \\ (1-x)^2 - \frac{1}{16}, & x \in (\frac{3}{4}, 1). \end{cases}$$

It is now easy to check that the third integral in the expression $\mu_{vvy}(\bar{t}z)$ (see (b_6)) is different from zero, while all the other integrals vanish. Hence, in the point $\bar{t}z$ condition (b_6) holds.

Next, by symmetry arguments it follows easily that

$$60\int_0^1 v^3 v_v dx + 90\int_0^1 \bar{t}zvv_v^2 dx = 0,$$

and hence the point $\bar{t}z$ is not a swallowtail singularity; i.e., (b_7) holds. Finally, we show that in $\bar{t}z$ condition (b_8) is verified; in particular, this implies that Φ is 1-1-1-1-transverse in $\bar{t}z$. First, note that

$$-v_{vv}'' = -6(v^3 + 3\bar{t}zv_v v)$$

and hence (b_8) can be written as

$$\begin{aligned} &270\int_0^1 v^2 v_v^2 dx - 10\int_0^1 (-v_{vv}'')v_{vv} dx + 90\int_0^1 \bar{t}zv_v^3 dx \\ &= 270\int_0^1 v_v^2 dx - 10\int_0^1 (v_{vv}')^2 dx + 90\int_0^1 \bar{t}zv_v^3 dx. \end{aligned}$$

By (41) we calculate

$$\int_0^1 v_v^2 dx = \frac{1}{5}, \quad \int_0^1 \bar{t}zv_v^3 dx = -\frac{3}{35}.$$

Furthermore, from

$$v_{vv}'' = \begin{cases} -3 + 36x^2, & 0 < x < \frac{1}{2}, \\ -3 + 36(1-x)^2, & \frac{1}{2} < x < 1 \end{cases}$$

we get

$$v_{vv} = \begin{cases} -\frac{3}{2}x^2 + 3x^4 + \frac{7}{80}, & 0 < x < \frac{1}{2}, \\ -\frac{3}{2}(1-x)^2 + 3(1-x)^4 + \frac{7}{80}, & \frac{1}{2} < x < 1, \end{cases}$$

and then $\int_0^1 (v_{vv}')^2 dx = \frac{1}{10}$. A calculation now yields (40).

The proof of the theorem is now completed by approximation: Let

$$z_n(x) := \begin{cases} (\cos \pi x)^{1/n}, & 0 \le x \le \frac{1}{2}, \\ -(-\cos \pi x)^{1/n}, & \frac{1}{2} \le x \le 1, \end{cases} \qquad n \in N.$$

Then $z_n \in E$ and $z_n \to z$ in $L^p(0,1)$ for all $p \ge 1$.

We claim that, for each $0 < \lambda < \lambda_2$, there exists a unique $t_n(\lambda) > 0$ such that $t_n(\lambda) z_n$ is a singular point; i.e., $t_n(\lambda) z_n \in S = S(\lambda)$. Indeed, using the variational characterization of $\mu(tz_n)$,

$$\mu(tz_n) = \inf_{\{w \in E, \int w^2 = 1\}} \int_0^1 (w'^2 - \lambda w^2 + 3t^2 z_n^2 w^2) dx,$$

we see that $\mu(0) = -\lambda$, $\mu(tz_n)$ increases in $t > 0$, and $\mu(tz_n) \to +\infty$ as $t \to +\infty$ for any fixed $n \in N$. Hence, there is a unique $t_n(\lambda)$ such that $\mu(t_n(\lambda) z_n) = 0$, and hence (b_1), $\forall n \in N$.

Denoting by $t(\lambda) = (\lambda/3)^{1/2}$ the unique t such that $\mu(tz) = 0$, we show that $(t_n(\lambda) z_n)^2 \to (t(\lambda) z)^2 = \lambda/3$ in $L^p(0,1), \forall p \ge 1$. Indeed, since $z(x) > z_{n+1}(x) > z_n(x)$, $\forall x \in (0,1), \forall n \in N$, we obtain $t_n(\lambda) > t_{n+1}(\lambda) > t(\lambda)$, $\forall n \in N$. Hence, $t_n(\lambda) \to t_0(\lambda) \ge t(\lambda)$ and then $(t_n(\lambda) z_n)^2 \to t_0^2(\lambda)$ in $L^p(0,1), \forall p \ge 1$. Denoting $v_n = v(t_n(\lambda) z_n)$, this implies

$$\begin{aligned} 0 &= \int_0^1 (v_n'^2 - \lambda v_n^2 + 3(t_n(\lambda) z_n)^2 v_n^2) dx \\ &\ge \int_0^1 (v_n'^2 - \lambda v_n^2 + 3t_0^2(\lambda) v_n^2) dx - 3 \int_0^1 |(t_n(\lambda) z_n)^2 - t_0^2(\lambda)| v_n^2 dx \\ &\ge \mu(t_0(\lambda) z) - 3 \int_0^1 |t_n^2(\lambda) z_n^2 - t_0^2(\lambda)| dx |v_n^2|_{L^\infty}; \end{aligned}$$

i.e., $\mu(t_0(\lambda) z) \le 0$. Since, on the other hand, $\mu(t_0(\lambda) z) \ge \mu(t(\lambda) z) = 0$, we conclude that $\mu(t_0(\lambda) z) = 0$, which yields $t_0(\lambda) = t(\lambda)$ by the monotonicity in t of $\mu(tz)$.

Now, denoting $t_n = t_n(\lambda)$, note that by

$$-v_n'' - \lambda v_n + 3(t_n z_n)^2 v_n = 0 = -v'' - \lambda v + 3(t(\lambda) z)^2 v$$

we get

$$-(v - v_n)'' - \lambda(v - v_n) + 3(t(\lambda) z)^2 (v - v_n) = 3((t_n z_n)^2 - (t(\lambda) z)^2) v_n.$$

Since the right-hand side converges to 0 in $L^2(0,1)$, one concludes that $v_n \to v$ in $C(0,1)$.

Furthermore, suppose that $p \in L^\infty(0,1)$ and $(p_n) \subset E$ with $p_n \to p$ in $L^p(0,1)$, $\forall p \ge 1$. Denote $v_{n,p_n} = v_n'(t_n z_n)[p_n]$ and $v_p := v'(t(\lambda) z)[p]$; we show that $v_{n,p_n} \to v_p$ in $C(0,1)$; in fact, taking the derivative of $-v_n'' - \lambda v_n + 3(t_n z_n)^2 v_n = 0$ in direction p_n we get

$$-v_{n,p_n}'' - \lambda v_{n,p_n} + 3(t_n z_n)^2 v_{n,p_n} = -6 t_n z_n v_n p_n$$

and, similarly,

$$-v_p'' - \lambda v_p + 3(t(\lambda) z)^2 v_p = -6 t(\lambda) z v p.$$

Subtracting the two equations yields

$$\begin{aligned}&-(v_{n,p_n}-v_p)''-\lambda(v_{n,p_n}-v_p)+3(t(\lambda)z)^2(v_{n,p_n}-v_p)\\&=-6t_nz_nv_np_n+6t(\lambda)zvp+3(t(\lambda)z)^2v_{n,p_n}-3(t_nz_n)^2v_{n,p_n};\end{aligned}$$

the right-hand side tends to 0 in $L^2(0,1)$, which yields that $v_{n,p_n}\to v_p$ in $C^1(0,1)$.

We next prove (b_5); i.e., we show that for each $n\in N$ we can choose a λ_n^* with $\lambda_n^*\to\lambda^*$ such that

$$\int_0^1 v_n^4dx+3\int_0^1 t_n(\lambda_n^*)z_nv_n^2v_{n,v_n}dx=0.$$

In fact, using $t_nz_nv_n^2=-\frac{1}{6}(-d^2/dx^2-\lambda+3(t_nz_n)^2)v_{n,v_n}$, we get

$$\begin{aligned}&\int_0^1 v_n^4dx+3\int_0^1 t_nz_nv_n^2v_{n,v_n}dx\\&=\int_0^1 v_n^4dx-\frac{1}{2}\int_0^1(v'_{n,v_n})^2dx-\frac{1}{2}\int_0^1(-\lambda+3(t_nz_n)^2)v_{n,v_n}^2dx\\&=\int_0^1 v_n^2dx-\frac{1}{2}2\int_0^{1/2}9t^2(\lambda)(2x)^2+\epsilon_n(\lambda)=1-\frac{\lambda}{2}+\epsilon_n(\lambda)\end{aligned}$$

with $\epsilon_n(\lambda)$ continuous in λ and $\epsilon_n\to 0$ uniformly in λ as $n\to+\infty$; here we have used the specific form of v_v given in (41). From this we see that, for every sufficiently large n, there is a λ_n^* as claimed, i.e., such that (b_5) is verified.

In what follows we assume that $\lambda=\lambda_n^*$. We then conclude that $t_n:=t_n(\lambda_n^*)\to\bar{t}=t(\lambda^*)$. Using the fact that $t_nz_n\to\bar{t}z$ in $L^p(0,1)$ and $v_n\to v$ in $C(0,1)$, one now concludes that t_nz_n is 1-transverse; i.e., (b_2) holds for n sufficiently large.

Furthermore, since t_nz_n is odd and $v_n=v(t_nz_n)$ even, we find that t_nz_n is not a fold; i.e., (b_3) is satisfied.

To verify the 1-1-transversality choose

$$w_n(x)=\begin{cases}z_n(2x), & 0<x<\frac{1}{2}\\ -z_n(2(x-\frac{1}{2})), & \frac{1}{2}<x<1.\end{cases}$$

By symmetry arguments one sees that $\int_0^1 t_nz_nv_n^2w_ndx=0$, i.e., $w_n\in T_{t_nz_n}S_1$. Furthermore, since $w_n\to w$ in $L^p(0,1),\forall p\geq 1$, and $v_n\to v$ in $C(0,1)$, $t_nz_n\to\bar{t}z$ in $L^p(0,1)$, and $v_{n,w_n}\to v_w$ in $C(0,1)$ (by taking $p_n=w_n$ in the formula above), we have that

$$\mu_{v_nw_n}(t_nz_n)=6\int_0^1 w_nv_n^3dx+18\int_0^1 t_nz_nv_n^2v_{n,w_n}^2dx\neq 0$$

for n sufficently large; hence (b_4).

To verify the 1-1-1-transversality set $y_n(x)=z_n(x)w_n(x)-\delta_ns(x)$, where

$$s(x):=\begin{cases}0\,, & x\in(0,\frac{1}{4})\cup(\frac{3}{4},1)\\ \sin^3(4\pi x)\,, & x\in[\frac{1}{4},\frac{3}{4}]\end{cases}$$

and $\delta_n\in R$ is such that $\int_0^1 t_nz_nv_n^2y_ndx=0$ (i.e., $y_n\in T_{t_nz_n}S_1$); clearly, $\delta_n\to 0$ and $y_n\to y$ (in L^p) as $n\to\infty$. Furthermore, since y_n is odd and v_n is even, and t_nz_n is odd and v_{n,y_n} is even (with respect to $x=\frac{1}{2}$), we have $\int_0^1 y_nv_n^3dx+$

$3\int_0^1 t_n z_n v_n^2 v_{n,y_n} dx = 0$, i.e., $y_n \in T_{t_n z_n} S_{1,1}$. And since $\mu_{v_n v_n y_n}(t_n z_n) \to \mu_{vvy}(\bar{t}z)$ (using the integral expressions), we have that (b_6) is verified for n sufficiently large.

Again by symmetry arguments, one sees that $t_n z_n$ is not a swallowtail singularity for n sufficiently large; i.e., (b_7) is satisfied.

Finally, verifying that $v_{n,v_n v_n} \to v_{vv}$ in $C^1(0,1)$ as well, we see that, for n sufficiently large, the butterfly condition (b_8) is satisfied. □

We mention that this result yields (locally) the following structure.

COROLLARY 21. *Under the conditions of Theorem* 20 *we have the following:*

(a) *Set* $\Sigma_1 = S_1 \cap U(t_n z_n)$, *where* $U \subset E$ *is a suitable neighbourhood of* $t_n z_n$ (*with* n *fixed sufficiently large*); *then*

Σ_1 *contains a smooth submanifold* Σ_2 *of codimension* 2 (*with respect to* E) *such that* $\Sigma_1 \backslash \Sigma_2$ *has exactly two components which consist of fold singularities*;

Σ_2 *contains a smooth submanifold* Σ_3 *of codimension* 3 (*with respect to* E) *such that* $\Sigma_2 \backslash \Sigma_3$ *has exactly two components which consist of cusp singularities*;

Σ_3 *contains a smooth submanifold* Σ_4 *of codimension* 4 (*with respect to* E) *such that* $\Sigma_3 \backslash \Sigma_4$ *has exactly two components which consist of swallowtail singularities*;

Σ_4 *consists of butterfly singularities.*

(b) *There exists an open region* $F_5 \subset F$ *such that, for* $h \in F_5$, *equation* (1) *has* (*locally*) *exactly five solutions.*

Proof. Statement (a) follows by Theorem 20 and the "stratified" structure of the sets $S_{1,1,1,1} \subset S_{1,1,1} \subset S_{1,1} \subset S_1$. Indeed, by the 1-transversality the set S_1 is locally a codimension 1 manifold Σ_1 in E containing (by the 1-1-transversality) a codimension 2 (with respect to E) submanifold Σ_2 of higher singularities. This means that $\Sigma_1 \backslash \Sigma_2$ consists of fold singularities. By (b_5) and (b_6), Σ_2 contains a codimension 3 (with respect to E) submanifold Σ_3 of singularities higher than cusps, and $\Sigma_3 \backslash \Sigma_2$ consists of cusp singularities. By (b_7) and (b_8), Σ_3 contains a codimension 4 submanifold Σ_4 consisting of butterfly singularities and $\Sigma_3 \backslash \Sigma_4$ consists of swallowtail singularities.

(b) follows from the structure of butterfly singularities; in fact, let β denote the normal form mapping for the butterfly singularity given by $\beta(t,s,r,q,x) = (t^5 + st^3 + rt^2 + qt, s, r, q, x)$, and suppose that the diffeomorphism $\alpha : U(t_n z_n) \subset E \to \alpha(U) \subset R^4 \times X$ maps $t_n z_n$ into the origin in $R^4 \times X$ (cf. Prop. 18). Choose $\bar{s} < 0$ and $\bar{t} > 0$ sufficiently close to zero such that $(t^5 + \bar{s}t^3 + \bar{q}t, \bar{s}, 0, \bar{q}, \mathbf{0}) = (0, \bar{s}, 0, \bar{q}, \mathbf{0})$ has five solutions in $\alpha(U)$. Then, for $\bar{h} = \gamma^{-1}(0, \bar{s}, 0, \bar{q}, \mathbf{0})$ equation (2) has five solutions, and the same holds for all h in the same component of $\bar{h}$ in $V \backslash \Phi(S)$, where $V = V(\Phi(\bar{u})) \subset F$; see Proposition 18. □

For a rectangle $\Omega = [0,a] \times [0,b]$, one obtains the existence of butterfly singularities with the same arguments and for the same range of parameter values λ as for the interval, taking into account the results in [19].

REFERENCES

[1] A. AMBROSETTI AND G. PRODI, *On the inversion of some differentiable mappings between Banach spaces*, Ann. Math. Pura Appl., 93 (1973), pp. 231–247.

[2] M. S. BERGER AND P. T. CHURCH, *Complete integrability and perturbation of a nonlinear Dirichlet problem* I, Indiana Univ. Math. J., 28 (1979), pp. 935–952.

[3] ———, *Complete integrability and perturbation of a nonlinear Dirichlet problem* II, Indiana Univ. Math. J., 29 (1980), pp. 715–735.

[4] M. S. BERGER, P. T. CHURCH, AND J. G. TIMOURIAN, *Folds and cusps in Banach spaces with applications to nonlinear partial differential equations*, Indiana Univ. Math. J., 34 (1985), pp. 1–19.

[5] V. CAFAGNA AND F. DONATI, *Un résultat global de multiplicité pour un problème differentiel non linéaire du premier ordre*, C. R. Acad. Sci. Paris Sér. I. Math., 300 (1985), pp. 523–526.

[6] ———, *Singularity theory of Fredholm maps and the number of solutions to some nonlinear first order differential equations*, preprint.

[7] V. CAFAGNA AND G. TARANTELLO, *Multiple solutions for some semilinear elliptic equations*, Math. Ann., 276 (1987), pp. 643–656.

[8] P. T. CHURCH AND J. G. TIMOURIAN, *The singular set of a non-linear elliptic operator*, Michigan Math. J., 35 (1988), pp. 197–213.

[9] P. T. CHURCH, T. N. DANCER, AND J. G. TIMOURIAN, *The structure of a nonlinear elliptic operator*, Trans. Amer. Math. Soc., 338 (1993), pp. 1–42.

[10] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 1, John Wiley, New York, 1989.

[11] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and their Singularities*, Springer, New York, 1973.

[12] M. GOLUBITSKY AND D. SCHAEFFER, *Singularities and Groups in Bifurcation Theory*, Vol. 1, in Appl. Math. Sci., 51, Springer, New York, 1985.

[13] M. GOLUBITSKY, I. STEWART, AND D. SCHAEFFER, *Singularities and Groups in Bifurcation Theory*, Vol. 2, in Appl. Math. Sci. 69, Springer, New York, 1988.

[14] F. LAZZERI AND A. M. MICHELETTI, *An application of singularity theory to nonlinear differentiable mappings between Banach spaces*, J. Nonlinear Anal., 11 (1986), pp. 795–808.

[15] H. P. MCKEAN AND J. C. SCOVEL, *Geometry of some simple nonlinear differential operator*, Ann. Scuola Norm. Sup. Pisa Cl. Sc. (4), 13 (1986), pp. 299–346.

[16] P. MICHOR, *The division theorem on Banach spaces*, Oesterr. Ak. Wiss., II-189-1-3 (1980), pp. 1–18.

[17] B. MORIN, *Formes canoniques des singularités d'une application différentiable*, Comptes Rendus Acad. Sci. Paris, 260 (1965), pp. 5662–5665 and 6503–6506.

[18] B. RUF, *Singularity theory and the geometry of a nonlinear elliptic equation*, Ann. Scuola Norm. Sup. Pisa Cl. Sc., (4), 17 (1990), pp. 1–33.

[19] ———, *Forced secondary bifurcation in an elliptic boundary value problem*, Differential Integral Equations, 5 (1992), pp. 793–804.

[20] R. THOM, *Les singularités des applications différentiables*, Ann. Inst. Fourier, 6 (1955–1956), pp. 43–87.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1361-1385, September 1995

015

CONSTRUCTION ET RÉGULARITÉ DES FONCTIONS D'ÉCHELLE*

LOÏC HERVE†

Résumé. Utilisant les propriétés spectrales d'opérateurs de la forme $P_w f(x) = w(\frac{x}{2})f(\frac{x}{2}) + w(\frac{x}{2}+\frac{1}{2})f(\frac{x}{2}+\frac{1}{2})$, nous étudions les fonctions d'échelle ϕ associées aux filtres d'échelle de longueur infinie. Nous calculons le coefficient de Sobolev de ϕ et plus généralement le plus grand coefficient s tel que $\int_{-\infty}^{+\infty} |\hat{\phi}(\lambda)|^p(1+|\lambda|^{ps})d\lambda < +\infty$, où $1 \leq p < +\infty$. Nous appliquons les résultats aux interpolations dyadiques continues.

Abstract. Using spectral properties of the operators $P_w f(x) = w(\frac{x}{2})f(\frac{x}{2}) + w(\frac{x}{2}+\frac{1}{2})f(\frac{x}{2}+\frac{1}{2})$, we study the scaling functions ϕ associated to filters with infinite length. We compute the Sobolev coefficient of ϕ and more generally the larger coefficient s such that $\int_{-\infty}^{+\infty} |\hat{\phi}(\lambda)|^p(1+|\lambda|^{ps})d\lambda < +\infty$, where $1 \leq p < +\infty$. We apply the results to dyadic interpolations.

Mots-clé. ondelette, analyse multirésolution, fonction d'échelle, opérateur de transfert, interpolation dyadique.

Key words. wavelet, multiresolution approximation, scaling function, transfer operator, dyadic interpolation.

AMS subject classifications. 39B32, 42C15, 47B07, 41A05.

1. Introduction: Analyses multirésolutions et filtres d'échelle.

Une analyse multirésolution [25], [23] est par définition une famille $(V_j)_{j\in\mathbf{Z}}$ de sous-espaces fermés de $L^2(\mathbf{R})$ tels que

(a) $\cap_{j\in\mathbf{Z}} V_j = \{\vec{0}\}$ et $\overline{\cup_{j\in\mathbf{Z}} V_j} = L^2(\mathbf{R})$.

(b) $V_j \subset V_{j+1}$.

(c) $f \in V_j$ est équivalent à $f(2^{-j}\cdot) \in V_0$).

(d) Il existe une fonction $g \in V_0$, appelée fonction d'échelle, telle que la famille $\{g(\cdot + k), k \in \mathbf{Z}\}$ forme une base de Riesz de V_0.

La condition (d) exprime que l'ensemble des combinaisons linéaires finies des fonctions $g(\cdot + k)$ est dense dans V_0, et qu'il existe une constante $c > 0$ telle que l'on ait, pour toute suite $(c_k)_{k\in\mathbf{Z}} \in \ell^2(\mathbf{Z})$,

$$\frac{1}{c}\sum_{k\in\mathbf{Z}} |c_k|^2 \leq \| \sum_{k\in\mathbf{Z}} c_k g(\cdot + k)\|^2_{L^2(\mathbf{R})} \leq c \sum_{k\in\mathbf{Z}} |c_k|^2. \tag{1}$$

Toute fonction f de V_0 s'écrit de manière unique sous la forme $f = \sum_{k\in\mathbf{Z}} c_k g(\cdot + k)$, où $(c_k)_{k\in\mathbf{Z}} \in \ell^2(\mathbf{Z})$. D'autre part, en vertu de (c), la famille $\{2^{j/2} g(2^j \cdot + k), k \in \mathbf{Z}\}$ forme une base de Riesz de V_j, pour tout $j \in \mathbf{Z}$. En particulier, puisque $g \in V_1$ (condition (b)), il existe une unique suite $(a_k)_{k\in\mathbf{Z}} \in \ell^2(\mathbf{Z})$ telle que

$$g(x) = \sum_{k\in\mathbf{Z}} a_k g(2x + k).$$

Par passage à la transformée de Fourier définie par

$$\hat{f}(\lambda) = \int_{-\infty}^{+\infty} f(x) e^{-2i\pi\lambda x} dx,$$

* Received by the editors November 9, 1992; accepted for publication (in revised form) January 4, 1994.

† I.R.M.A.R., Université de Rennes 1, Laboratoire de probabilités, Campus de Beaulieu, 35042 Rennes cedex, France.

l'équation ci-dessus devient $\hat{g}(\lambda) = H(\frac{\lambda}{2})\hat{g}(\frac{\lambda}{2})$, où H est la série trigonométrique de coefficients $(\frac{a_k}{2})_{k\in\mathbf{Z}}$, appelée filtre d'échelle associé à g. Dans les exemples d'analyses multirésolutions utilisées en pratique, H est toujours de classe $\mathcal{C}^\infty$, souvent polynomiale, et telle que $H(0) = 1$, $H(\frac{1}{2}) = 0$.

Réciproquement, considérons une suite $(h_k)_{k\in\mathbf{Z}} \in \ell^2(\mathbf{Z})$ quelconque telle que la série trigonométrique

$$m_0(\lambda) = \frac{1}{2}\sum_{k\in\mathbf{Z}} h_k e^{2i\pi k\lambda}$$

soit de classe $\mathcal{C}^\infty$ et telle que $m_0(0) = 1$, $m_0(\frac{1}{2}) = 0$. On pose, pour tout $\lambda \in \mathbf{R}$,

$$\hat{\phi}(\lambda) = \prod_{k=1}^{+\infty} m_0\left(\frac{\lambda}{2^k}\right). \tag{2}$$

La fonction $\hat{\phi}$ est définie, continue sur $\mathbf{R}$ [25], [8], et vérifie $\hat{\phi}(0) = 1$ et l'équation

$$\hat{\phi}(\lambda) = m_0\left(\frac{\lambda}{2}\right)\hat{\phi}\left(\frac{\lambda}{2}\right). \tag{3}$$

L'objet de ce travail est de déterminer des conditions nécessaires et suffisantes sur $m_0 \in \mathcal{C}^\infty$ pour que l'on ait (successivement) les deux propriétés suivantes.

(P1) La fonction $\hat{\phi}$, et donc sa transformée de Fourier inverse ϕ, sont de carré intégrable sur $\mathbf{R}$;

(P2) la famille $\{\phi(\cdot + k), k \in \mathbf{Z}\}$ forme un système de Riesz (i.e., vérifie (1)).

Le lien avec les analyses multirésolutions est donné par la propriété suivante.

Sous les conditions (P1), (P2), *la famille* $(V_j)_{j\in\mathbf{Z}}$, *où* V_j *est le sous-espace de* $L^2(\mathbf{R})$ *engendré par le système* $\{2^{\frac{j}{2}}\phi(2^j \cdot +k), k \in \mathbf{Z}\}$, *forme une analyse multirésolution* (*avec* $g = \phi$ *et* $H = m_0$).

En effet, les points (c) et (d) de la définition sont évidents. En outre, par passage à la transformée de Fourier inverse dans (3), la fonction ϕ satisfait à l'équation d'échelle

$$\phi(x) = \sum_{k\in\mathbf{Z}} h_k \phi(2x + k), \tag{4}$$

ce qui prouve (b). Pour établir les conditions asymptotiques (a), on pourra utiliser les résultats de [8, pp. 141–142]. En accord avec la terminologie usuelle rappelée plus haut, toute fonction m_0 satisfaisant à (P1) et (P2) sera appelée filtre d'échelle, ϕ étant alors la fonction d'échelle associée.

L'étude de la régularité de ϕ tient une place importante dans la théorie des analyses multirésolutions et ses applications. A cet effet, il existe deux approches bien distinctes: l'une où l'on calcule par une méthode directe (n'utilisant pas la transformée de Fourier) le coefficient d'Hölder optimal de ϕ (voir [9], [26], [28]); la seconde (fréquentielle) où l'on estime le coefficient de Sobolev de ϕ. Nous nous proposons ici de généraliser la méthode fréquentielle, et plus précisément de résondre le problème suivant.

(P3) Trouver des conditions pour que $\hat{\phi} \in L^p(\mathbf{R})$, où p est un réel quelconque tel que $1 \le p < +\infty$, et de calculer dans ce cas le coefficient

$$s_p = \sup\left\{s > 0 \ : \int_{-\infty}^{+\infty} |\hat{\phi}(\lambda)|^p (1 + |\lambda|^{ps}) d\lambda < +\infty\right\}.$$

Les cas $p = 1$ et 2 sont particulièrement intéressants pour estimer la régularité de ϕ. En effet, si $\phi \in L^1(\mathbf{R})$ et $s_1 > 0$ (respectivement, si $s_2 > \frac{1}{2}$), alors ϕ est hölderienne d'ordre $s_1 - \epsilon$ (respectivement, $s_2 - \frac{1}{2} - \epsilon$) pour tout $\epsilon > 0$.

Dans le cas particulier des filtres polynomiaux (la suite $(h_k)_{k\in\mathbf{Z}}$ est à support fini), une solution simple de (P1), (P2) a été donnée indépendamment dans [27], [20], et on trouvera le calcul de s_2 dans [27], [13], [20], et celui de s_1 pour les filtres polynomiaux positifs dans [5] (s_1 est alors le coefficient d'Hölder optimal de ϕ). Dans les articles précédemment cités, d'une part seul le cas polynomial est traité, et d'autre part, les études (directes ou fréquentielles) de la régularité de ϕ font appel à des calculs dont la complexité augmente avec la longueur du filtre m_0 (nombre de h_k non nuls).

Comme l'illustre l'énoncé ci-dessous (démontré dans le §4), l'intérêt de ce travail est de fournir une solution de (P1)–(P3) applicable aux filtres non polynomiaux, et telle que la complexité des calculs mis en jeu ne dépende pas, ou très peu, de la taille des filtres (dans le cas polynomial).

Supposons que m_0 admette un nombre fini $q+1$ de zéros, et que

$$m_0(\lambda) = \left(\frac{1+e^{2i\pi\lambda}}{2}\right)^r v(\lambda),$$

avec $r \in \mathbf{N}^*$ et $v(\frac{1}{2}) \neq 0$ (la factorisation ci-dessus est naturelle puisque m_0 s'annule en $\frac{1}{2}$). On pose, pour tout $n \geq 1$,

$$S_n = \sum_{k=0}^{2^n-1} |v(\tfrac{k}{2})|^2 \cdots \left|v\left(\frac{k}{2^n}\right)\right|^2.$$

Pour simplifier, commençons par supposer que $q \leq 1$. Alors,

(i) *La suite $(S_{n+1}/S_n)_{n\geq 1}$ converge avec une vitesse exponentielle vers un réel noté β_2.*

(ii) *Une condition nécessaire et suffisante pour que m_0 soit un filtre d'échelle est que $|m_0(\frac{\lambda}{2})| + |m_0(\frac{\lambda}{2}+\frac{1}{2})| > 0$ pour tout $\lambda \in \mathbf{R}$, et que $\beta_2 < 2^{2r}$. Dans ce cas, on a*

$$s_2 = r - \frac{1}{2}\log_2 \beta_2.$$

Si $q \geq 2$, il faut ajouter, des conditions très simples sur les zéros de m_0 (par exemple, si $q = 2$, $|m_0(\frac{1}{6})| + |m_0(\frac{5}{6})| \neq 0$). D'autre part, le calcul de s_p pour p quelconque est identique (remplacer $|v|^2$ par $|v|^p$ dans la définition de S_n; voir §6). On montre dans [14] que $s_2 = r - \frac{1}{2}\log_2 \rho_a$, où ρ_a est le rayon spectral de l'opérateur P_w (défini ci-dessous) avec $w = |v|^2$. Sous les hypothèses précédentes, on a $\rho_a = \beta_2$; voir §4.

Dans le §2, nous présentons une première caractérisation des filtres d'échelle de classe $\mathcal{C}^\infty$. Cette partie permet d'introduire les deux principaux outils de ce travail: les compacts invariants et les opérateurs de tranfert P_w de la forme

$$P_w f(x) = w\left(\frac{x}{2}\right) f\left(\frac{x}{2}\right) + w\left(\frac{x}{2}+\frac{1}{2}\right) f\left(\frac{x}{2}+\frac{1}{2}\right), \qquad x \in [0,1]$$

où w est une fonction 1-périodique donnée, et f une fonction quelconque définie sur $[0,1]$. Ces deux dernières notions, classiques en théorie ergodique, ont été introduites dans le cadre des orthogonales par Conze et Raugi [7]. L'étude spectale de P_w (voir

§3), est basée sur un résultat récent de Hennion [15] qui complète le théorème classique de Ionescu-Tulcea et Marinescu [21]. Nous présentons ensuite une deuxième solution pour (P1), (P2), tout d'abord sous l'hypothèse $m_0 \in C^\infty$ (§4), puis dans le cas particulier des filtres polynomiaux (§5). Dans le §6, nous étudions (P3), et nous présentons un calcul numérique des coefficients s_1 et s_2 pour les filtres de Butterworth (non polynomiaux). Enfin nous appliquons les résultats précédents aux interpolations dyadiques continues (§7). Signalons enfin que les principaux résultats de ce papier (caractérisation des filtres d'échelle et calcul des s_p) ont été énoncés dans [17], et démontrés dans [20] dans le cas plus général où m_0 est hölderienne.

2. Filtres d'échelle: premiers critères. Dans ce paragraphe, m_0 est une fonction 1-périodique, à valeurs dans $\mathbf{C}$, de classe C^∞, telle que $m_0(0) = 1$. Cependant la plupart des résultats restent vérifiés si m_0 est seulement hölderienne. On note $u = |m_0|^2$, et on considère la fonction $\hat{\phi}$ définie par (2). Rappelons que $\hat{\phi}$ est continue sur $\mathbf{R}$. Si $\hat{\phi} \in L^2(\mathbf{R})$, on note ϕ sa transformée de Fourier inverse, qui est alors elle-même de carré intégrable. Dans toute la suite, nous désignerons par θ_ϕ la fonction 1-périodique, a priori à valeurs dans $[0, +\infty]$, donnée par

$$\theta_\phi(x) = \sum_{k \in \mathbf{Z}} |\hat{\phi}(x+k)|^2. \tag{5}$$

La fonction θ_ϕ est semi-continue inférieurement sur $\mathbf{R}$, comme limite croissante de fonctions continues. Si $\hat{\phi} \in L^2(\mathbf{R})$, alors $\theta_\phi \in L^1([0,1])$, et plus précisément, θ_ϕ est indéfiniment dérivable [25]. La propriété importante ici est la continuité de θ_ϕ.

2.1. Première résolution de (P1). Nous obtenons, grâce à (3),

$$\theta_\phi(x) = \sum_k \left| m_0\left(\frac{x}{2} + \frac{k}{2}\right)\right|^2 \left|\hat{\phi}\left(\frac{x}{2} + \frac{k}{2}\right)\right|^2,$$

soit encore, en séparant cette dernière somme selon les indices pairs et impairs,

$$\theta_\phi(x) = \left|m_0\left(\frac{x}{2}\right)\right|^2 \theta_\phi\left(\frac{x}{2}\right) + \left|m_0\left(\frac{x}{2} + \frac{1}{2}\right)\right|^2 \theta_\phi\left(\frac{x}{2} + \frac{1}{2}\right). \tag{6}$$

L'identité (6) exprime que θ_ϕ est invariante sous l'action de l'opérateur de transfert P_u défini par

$$P_u f(x) = u\left(\frac{x}{2}\right) f\left(\frac{x}{2}\right) + u\left(\frac{x}{2} + \frac{1}{2}\right) f\left(\frac{x}{2} + \frac{1}{2}\right), \qquad x \in [0,1], \tag{7}$$

où $u = |m_0|^2$, et où f est une fonction quelconque définie sur $[0,1]$, à valeurs dans $\mathbf{C}$.

THÉORÈME 2.1. *Soit m_0 une fonction 1-périodique, de classe C^∞, telle que $m_0(0) = 1$, et soit $\hat{\phi}$ définie par (2). Alors $\hat{\phi} \in L^2(\mathbf{R})$ si, et seulement si, il existe une fonction γ P_u-invariante, 1-périodique, positive ou nulle, de classe C^∞, et enfin non nulle en 0.*

Remarque. Si m_0 est seulement hölderienne, la formule (2) définit encore une fonction $\hat{\phi}$ continue sur $\mathbf{R}$, ce qui asssure la semi-continuité de θ_ϕ. Dans ce cas, $\hat{\phi} \in L^2(\mathbf{R})$ si, et seulement si, il existe une fonction γ P_u-invariante, 1-périodique, positive ou nulle, intégrable et semi-continue inférieurement sur $[0,1]$, et enfin non nulle en 0 (voir [20]).

En particulier, pour tout réel p tel que $1 \leq p < +\infty$, la fonction $|m_0|^{p/2}$ est hölderienne, et

$$|\hat{\phi}(\lambda)|^{p/2} = \prod_{k=1}^{+\infty} |m_0(\frac{\lambda}{2^k})|^{p/2}.$$

Appliquant la remarque précédente avec $|m_0|^{p/2}$ à la place de m_0, nous obtenons le corollaire suivant.

COROLLAIRE 2.2. *Soit m_0 une fonction 1-périodique, de classe $\mathcal{C}^\infty$, telle que $m_0(0) = 1$, et soit $\hat{\phi}$ définie par (2). Soient en outre p un réel tel que $1 \leq p < +\infty$, $u_p(x) = |m_0(x)|^p$, et enfin P_{u_p} l'opérateur défini comme dans (7), mais avec u_p à la place de u. Alors une condition nécessaire et suffisante pour que $\hat{\phi} \in L^p(\mathbf{R})$ est qu'il existe une fonction γ P_{u_p}-invariante, 1-périodique, positive ou nulle, intégrable et semi-continue inférieurement sur $[0,1]$, et enfin non nulle en 0.*

Démonstration du théorème 2.1. Si $\hat{\phi} \in L^2(\mathbf{R})$, alors $\gamma = \theta_\phi$ convient. Réciproquement, soit γ une fonction satisfaisant aux hypothèses de l'énoncé. Nous utiliserons le lemme suivant.

LEMME 2.3. *Soit w une fonction continue, 1-périodique quelconque. On a, pour toute fonction f continue, 1-périodique, et tout entier $n \geq 1$,*

$$\int_{-2^{n-1}}^{2^{n-1}} f\left(\frac{\lambda}{2^n}\right) \prod_{k=1}^{n} w\left(\frac{\lambda}{2^k}\right) d\lambda = \int_0^1 P_w^n f(\lambda) d\lambda.$$

Preuve du lemme. On a

$$\int_0^1 P_w f(\lambda) d\lambda = \int_0^1 w\left(\frac{\lambda}{2}\right) f\left(\frac{\lambda}{2}\right) d\lambda + \int_0^1 w\left(\frac{\lambda}{2}+\frac{1}{2}\right) f\left(\frac{\lambda}{2}+\frac{1}{2}\right) d\lambda.$$

La formule du lemme, pour $n = 1$, découle du changement de variables $\lambda' = \lambda + 1$ dans la dernière intégrale, et de la périodicité de w et f.

Pour $n \geq 2$, on procède par récurrence: supposons le lemme vérifié pour un entier $n \geq 1$ donné. On a

$$\begin{aligned}
\int_0^1 P_w^{n+1} f(\lambda) d\lambda &= \int_{-2^{n-1}}^{2^{n-1}} P_w f\left(\frac{\lambda}{2^n}\right) \prod_{k=1}^{n} w\left(\frac{\lambda}{2^k}\right) d\lambda \\
&= 2^n \int_{-\frac{1}{2}}^{\frac{1}{2}} P_w f(\lambda) \prod_{k=1}^{n} w(2^{n-k}\lambda) d\lambda \\
&= 2^n \Big[\int_{-\frac{1}{2}}^{\frac{1}{2}} w\left(\frac{\lambda}{2}\right) f\left(\frac{\lambda}{2}\right) \prod_{k=1}^{n} w(2^{n-k}\lambda) d\lambda \\
&\quad + \int_{-\frac{1}{2}}^{\frac{1}{2}} w\left(\frac{\lambda}{2}+\frac{1}{2}\right) f\left(\frac{\lambda}{2}+\frac{1}{2}\right) \prod_{k=1}^{n} w(2^{n-k}\lambda) d\lambda \Big].
\end{aligned}$$

Utilisant les changements de variables $\lambda' = \frac{\lambda}{2}$ et $\lambda' = \frac{\lambda}{2} + \frac{1}{2}$, respectivement, dans la première et la dernière des deux intégrales ci-dessus, on obtient, grâce à la périodicité de w et f,

$$\int_0^1 P_w^{n+1} f(\lambda) d\lambda = 2^{n+1} \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\lambda) \prod_{k=1}^{n+1} w(2^{n+1-k}\lambda) d\lambda,$$

et l'on conclut grâce au changement de variables $\lambda' = 2^{n+1}\lambda$. Le lemme est ainsi démontré.

Le lemme appliqué avec $f = \gamma$ et $w = u$ donne

$$\int_{-2^{n-1}}^{2^{n-1}} \gamma(\frac{\lambda}{2^n}) \prod_{k=1}^{n} u\left(\frac{\lambda}{2^k}\right) d\lambda = \int_0^1 P_u^n \gamma(\lambda) d\lambda = \int_0^1 \gamma(\lambda) d\lambda,$$

d'où, grâce au lemme de Fatou, $\gamma(0) \int_{-\infty}^{+\infty} |\hat{\phi}(\lambda)|^2 d\lambda \leq \int_0^1 \gamma(\lambda) d\lambda$, ce qui prouve bien que $\hat{\phi}$ est de carré intégrable sur $\mathbf{R}$. □

2.2. Première caractérisation des filtres d'échelle de classe $\mathcal{C}^\infty$. Rappelons que m_0 est un filtre d'échelle si les conditions (P1) et (P2) du §1 sont satisfaites. Supposons dans un premier temps que $\hat{\phi} \in L^2(\mathbf{R})$. On sait que θ_ϕ est continue, et il est prouvé dans [25], [8] que la famille $\{\phi(\cdot + k), k \in \mathbf{Z}\}$ forme un système de Riesz si, et seulement si, il existe une constante $c > 0$ telle que l'on ait, pour tout $x \in \mathbf{R}$,

$$\frac{1}{c} \leq \theta_\phi(x) \leq c. \tag{8}$$

Autrement dit, (P2) est équivalent à (8). On déduit facilement de (6) et (8) qu'une condition nécessaire pour que m_0 soit un filtre d'échelle est que

$$|m_0(\frac{x}{2})|^2 + |m_0(\frac{x}{2} + \frac{1}{2})|^2 > 0, \quad \text{pour tout } x \in \mathbf{R}. \tag{9}$$

D'autre part, comme $\hat{\phi}(0) = 1$, on a $\theta_\phi(0) \geq 1$. Les conditions (8) et (6) (avec $x = 0$) entraînent que $m_0(\frac{1}{2}) = 0$. Cette dernière condition est donc nécessaire pour (P1) et (P2).

Pour la résolution de (P2), nous utiliserons la notion de compact invariant [7]. Dans le cadre des filtres QMF et biorthogonaux, on trouvera une autre solution de (P2) dans [3] et [4].

Définition des compacts invariants. Soient S_0 et S_1 les applications de $[0,1]$ dans lui-même, définies par

$$S_0 x = \frac{x}{2} \quad \text{et} \quad S_1 x = \frac{x}{2} + \frac{1}{2}, \qquad x \in [0,1].$$

Soit w une fonction 1-périodique, continue, telle que

$$\left|w\left(\frac{x}{2}\right)\right| + \left|w\left(\frac{x}{2} + \frac{1}{2}\right)\right| > 0, \quad \text{pour tout } x \in \mathbf{R}. \tag{10}$$

Un compact K de $[0,1]$ est dit invariant pour w si, pout tout $x \in K$ et tout $\sigma \in \{S_0, S_1\}$ tels que $w(\sigma x) \neq 0$, on a $\sigma x \in K$. Le lien avec le problème (P2) est donné par la proposition suivante.

PROPOSITION 2.4. *Soit m_0 une fonction 1-périodique, de classe $\mathcal{C}^\infty$, vérifiant $m_0(0) = 1$ et (9). Alors l'ensemble $Z(\theta_\phi)$ des zéros de θ_ϕ sur $[0,1]$ est un compact invariant pour m_0, ne contenant ni 0 ni 1. En outre, si K est un compact invariant pour $|m_0|$, disjoint de $\{0,1\}$, alors θ_ϕ est identiquement nulle sur K.*

Démonstration de la proposition 2.4. Le fait que $Z(\theta_\phi)$ soit un compact invariant pour m_0 est évident d'après (6). On a en outre $\theta_\phi(0) = \theta_\phi(1) \geq |\hat{\phi}(0)|^2 = 1$, ce qui prouve la première assertion de la proposition.

Soient K un compact invariant pour m_0, disjoint de $\{0,1\}$, x un élément de K, et enfin $k \in \mathbf{Z}$ quelconque. Pour tout $y \in \mathbf{R}$, nous notons $[y]$ sa partie entière, et $\{y\} = y - [y]$ sa partie fractionnaire. Nous devons prouver que $\hat{\phi}(x+k) = 0$, ou encore, puisque m_0 est 1-périodique, que

$$m_0\left(\left\{\frac{x+k}{2}\right\}\right) m_0\left(\left\{\frac{x+k}{4}\right\}\right) \cdots m_0\left(\left\{\frac{x+k}{2^n}\right\}\right) \cdots = 0.$$

Or, pour tout $y \in \mathbf{R}$, on a $\{\frac{y}{2}\} = \frac{\{y\}}{2}$ ou $\{\frac{y}{2}\} = \frac{\{y\}}{2} + \frac{1}{2}$. Il existe donc une suite $(\sigma_n)_{n\geq 1}$ d'éléments de $\{S_0, S_1\}$ tels que $\{(x+k)/2^n\} = \sigma_n \ldots \sigma_1(x)$ pour chaque $n \geq 1$. Comme K est invariant, si l'on avait $m_0(\{(x+k)/2^n\}) \neq 0$, pour tout $n \geq 1$, alors $(\{(x+k)/2^n\})_{n\geq 1}$ serait une suite d'éléments de K convergeant vers 0, ce qui est impossible par hypothèse. Il existe donc un entier $n \geq 1$ tel que $m_0(\{(x+k)/2^n\}) = 0$, d'où $\hat{\phi}(x+k) = 0$. □

Remarque. La proposition précédente est encore valable si m_0 est uniformément hölderienne. En effet, il suffit de vérifier que $Z(\theta_\phi)$ est un compact, l'invariance étant à nouveau assurée par (6). Soit $(x_n)_{n\geq 1}$ une suite d'éléments de $Z(\theta_\phi)$ convergeant vers $x \in [0,1]$. Comme θ_ϕ est semi-continue inférieurement, nous avons $\theta_\phi(x) \leq \liminf_n \theta_\phi(x_n)$, d'où $\theta_\phi(x) = 0$. Ainsi, $Z(\theta_\phi)$ est un fermé de [0,1], et donc un compact.

Donnons maintenant une première caractérisation des filtres d'échelle de classe $\mathcal{C}^\infty$.

THÉORÈME 2.5. *Soit m_0 une fonction 1-périodique, de classe $\mathcal{C}^\infty$, telle que $m_0(0) = 1$, et soit $u = |m_0|^2$. Une condition nécessaire et suffisante pour que m_0 soit un filtre d'échelle est que m_0 vérifie (9), $m_0(\frac{1}{2}) = 0$, puis que P_u possède une fonction γ invariante, 1-périodique, de classe $\mathcal{C}^\infty$, stictement positive, et enfin que tout compact invariant pour m_0 contienne 0 ou 1.*

Dans ce cas, θ_ϕ est l'unique fonction P_u-invariante, continue et 1-périodique (à un scalaire multiplicatif près).

Remarque. La condition $m_0(\frac{1}{2}) = 0$ entraîne que $\hat{\phi}(k) = 0$ pour tout $k \in \mathbf{Z}$, $k \neq 0$, d'où $\theta_\phi(0) = 1$. En particulier, si m_0 est un filtre d'échelle, toute fonction h P_u-invariante, continue et 1-périodique, s'écrit $h = h(0)\theta_\phi$.

Démonstration du théorème 2.5. Grâce à (6) et (8), si m_0 est un filtre d'échelle, la fonction θ_ϕ, qui est de classe $\mathcal{C}^\infty$, est P_u-invariante, 1-périodique, et strictement positive, ce qui implique (9) et $m_0(\frac{1}{2}) = 0$. La propriété sur les compacts invariants résulte de la proposition précédente.

Réciproquement, supposons l'existence de la fonction γ de l'énoncé, et que tout compact invariant pour m_0 contienne 0 ou 1. En vertu du théorème précédent, $\hat{\phi}$, et donc ϕ, appartiennent à $L^2(\mathbf{R})$. On sait que θ_ϕ est alors de classe $\mathcal{C}^\infty$, et, d'après la proposition 2.4, que l'ensemble $Z(\theta_\phi)$ des zéros de θ_ϕ sur $[0,1]$ est un compact invariant pour m_0, ne contenant ni 0 ni 1. On en déduit que $Z(\theta_\phi)$ est vide, et donc que θ_ϕ vérifie (8), ce qui prouve que m_0 est un filtre d'échelle.

Pour prouver la dernière propriété du théorème, on considère la fonction $\tilde{u}(x) = (\theta_\phi(2x))^{-1}\theta_\phi(x)u(x)$. Désignant par $P_{\tilde{u}}$ l'opérateur défini par (7) avec $\tilde{u}$ à la place de u, on obtient $P_{\tilde{u}}f = \theta_\phi^{-1}P_u(\theta_\phi f)$, d'où $P_{\tilde{u}}1 = 1$, ou encore $\tilde{u}(\frac{x}{2}) + \tilde{u}(\frac{x}{2} + \frac{1}{2}) = 1$ pour tout $x \in \mathbf{R}$. La fonction $\tilde{u}$ possède les mêmes compacts invariants que u_0. En outre, si g est une fonction P_u-invariante, continue et 1-périodique, alors $P_{\tilde{u}}(\theta_\phi^{-1}g) = \theta_\phi^{-1}g$. Il reste donc à prouver que, si $\tilde{g}$ est une fonction $P_{\tilde{u}}$-invariante, continue et 1-périodique, alors $\tilde{g}$ est identiquement constante. Soient $K_0 = \{x \in [0,1] : \tilde{g}(x) = \inf \tilde{g}\}$ et $K_1 = \{x \in [0,1] : \tilde{g}(x) = \sup \tilde{g}\}$. On montre facilement que K_0 et K_1 sont

des compacts invariants pour $\tilde{u}$. Ils contiennent donc 0 ou 1, d'où $\inf \tilde{g} = \sup \tilde{g} = \tilde{g}(0) = \tilde{g}(1)$. Le théorème est finalement démontré. □

2.3. Cas où m_0 a un nombre fini de zéros: Description des compacts invariants. La condition du théorème 2.5 sur les compacts invariants, difficile à vérifier dans la pratique, peut être simplifiée si m_0 possède un nombre fini de zéros sur [0,1]. A cet effet, nous utiliserons les définitions suivantes.

Soit $m \in \mathbf{N}^*$. Nous dirons que $x \in [0,1]$ est *un point périodique d'ordre* m s'il existe m éléments $\sigma_1, \ldots, \sigma_m$ de $\{S_0, S_1\}$ tels que $\sigma_m \cdots \sigma_1 x = x$, et si m est le plus petit entier pour lequel on a une telle relation. La famille $\{\sigma_1, \ldots, \sigma_m\}$ vérifiant la relation ci-dessus est unique, et l'on pose

$$\mathcal{C}_x = \{\sigma_k \ldots \sigma_1 x,\ k = 1, \ldots, m\}.$$

Nous appelons *cycle périodique d'ordre* m tout sous-ensemble fini C de $[0,1]$ de la forme $\mathcal{C}_x$, où x est un point périodique d'ordre m. Notons qu'un cycle périodique C est invariant pour une fonction w si l'on a $w(y + \frac{1}{2}) = 0$ pour chaque élément y de C. Les propriétés ci-dessous sont démontrées dans [18].

• *Les points périodiques d'ordre inférieur ou égal à* q, $q \in \mathbf{N}^*$, *sont de la forme* $k/(2^p - 1)$, *où* $p \in \{1, \ldots, q\}$ *et* $k \in \{0, 1, \ldots, 2^p - 1\}$.

• *Soit* w *une fonction 1-périodique, continue, positive ou nulle, vérifiant* (10) *et possédant un nombre fini de zéros. En outre soit* P_w *l'opérateur défini par* (7) *avec* w *au lieu de* u. *Si* γ *est une fonction continue sur* $[0,1]$, *positive ou nulle, telle que* $P_w\gamma = \beta\gamma$, *où* $\beta > 0$, *alors l'ensemble* $Z(\gamma)$ *des zéros de* γ *est, ou bien vide, ou bien une réunion finie de cycles périodiques invariants pour* w.

• *Soit* m_0 *une fonction 1-périodique, continue, vérifiant* (9), $m_0(0) = 1$, $m_0(\frac{1}{2}) = 0$, *et admettant* $q + 1$ *zéros. Les trois propriétés suivantes sont équivalentes.*

(i) Tout compact invariant pour m_0 *contient* 0 *ou* 1.

(ii) Il n'existe pas de cycles périodiques invariants pour m_0 *autres que* $\{0\}$ *et* $\{1\}$.

(iii) Les points périodiques x, *différents de 0 et 1, d'ordre inférieur ou égal à* q, *vérifient la condition suivante:*

$$\exists y \in \mathcal{C}_x \quad \text{tel que } m_0(y + \tfrac{1}{2}) \neq 0. \tag{11}$$

Notons que (11) nécessite un nombre fini de vérifications élémentaires. La condition (11) est toujours satisfaite si $q \leq 1$. Pour $q \geq 2$, il suffit de vérifier que m_0 n'est pas identiquement nulle sur un certain nombre de sous-ensembles finis de $[0,1]$ (dépendants uniquement de q), par exemple $\{\frac{1}{6}, \frac{5}{6}\}$ si $q = 2$, ou encore sur $\{\frac{1}{6}, \frac{5}{6}\}$ et $\{\frac{1}{14}, \frac{11}{14}, \frac{9}{14}\}$ si $q = 3$.

3. Description spectrale des opérateurs de transfert. Soit $(E, \|\ \|_\infty)$ l'espace des fonctions continues sur $[0,1]$, à valeurs dans $\mathbf{C}$, muni de la norme uniforme. Soit w une fonction de E à valeurs positives ou nulles. Nous désignerons par P_w l'opérateur défini sur E par

$$P_w f(x) = w\left(\frac{x}{2}\right) f\left(\frac{x}{2}\right) + w\left(\frac{x}{2} + \frac{1}{2}\right) f\left(\frac{x}{2} + \frac{1}{2}\right), \qquad x \in [0,1]. \tag{12}$$

Il est clair que l'opérateur P_w est borné sur E, et qu'il est positif (il conserve le sous-ensemble de E formé des fonctions positives). Une première conséquence de la positivité de P_w est que $\|P_w^n f\|_\infty \leq \|P_w^n 1\|_\infty \|f\|_\infty$ pour tout $n \geq 1$ et tout $f \in E$.

Soit ρ_w le rayon spectral de P_w sur E. On a

$$\rho_w = \lim_{n \to +\infty} ||P_w^n 1||_\infty^{1/n}. \tag{13}$$

L'étude spectrale de P_w, faite ci-dessous pour w hölderienne, s'appuie sur la propriété de quasi-compacité [15] dont nous rappelons la définition.

DÉFINITION. *Soit $(L, |||\cdot|||_L)$ un espace de Banach complexe. Un opérateur T borné sur L, de rayon spectral $\rho(T)$, est dit quasi-compact s'il existe un réel r tel que $0 \leq r < \rho(T)$, et deux sous-espaces N et F supplémentaires dans L, stables par T, tels que*

$1 \leq \dim N < +\infty$ et $T_{|N}$ n'a que des valeurs propres de module $\geq r$,

F est fermé et le rayon spectral de $T_{|F}$ est strictement inférieur à r.

Pour tout réel $\alpha, 0 < \alpha \leq 1$, nous notons E^α le sous-espace de E constitué des fonctions vérifiant

$$m_\alpha(f) = \sup\left\{\frac{|f(x) - f(y)|}{|x-y|^\alpha}, x, y \in [0,1], x \neq y\right\} < +\infty.$$

Nous munissons E^α de la norme $|||\ |||_\alpha$ définie par

$$|||f|||_\alpha = m_\alpha(f) + ||f||_\infty.$$

Il est clair que P_w est un opérateur borné sur E^α. Soit λ une valeur propre quelconque de P_w sur E^α. Si, pour un entier $i \geq 1$, les espaces $\mathrm{Ker}(P_w - \lambda)^i$ et $\mathrm{Ker}(P_w - \lambda)^{i+1}$ coïncident, nous noterons

$$\nu(\lambda) = \inf\{i \in \mathbf{N}^* : \mathrm{Ker}(P_w - \lambda)^i = \mathrm{Ker}(P_w - \lambda)^{i+1}\},$$

l'ordre, ou encore l'indice (cf. [12]), de λ pour P_w sur E^α. Le théorème suivant est démontré dans [15], [18].

THÉORÈME 3.1. *Soit w une fonction de E^α, positive ou nulle sur $[0,1]$. Sur l'espace E^α, P_w est quasi-compact, admet ρ_w comme rayon spectral et valeur propre, et enfin possède une fonction propre associée γ positive ou nulle. Les valeurs spectrales de module ρ_w sont en nombre fini et constituent des valeurs propres de P_w. Toute valeur propre λ de module ρ_w est telle que $\nu(\lambda) < +\infty$, $\dim \mathrm{Ker}(P_w - \lambda)^{\nu(\lambda)} < +\infty$, et l'on a plus précisément*

$$\nu(\lambda) \leq \nu(\rho_w) < +\infty.$$

On dispose en outre de la décomposition suivante

$$E^\alpha = \left(\oplus_{|\lambda|=\rho_w} \mathrm{Ker}(P_w - \lambda)^{\nu(\lambda)}\right) \oplus F,$$

où F est un sous-espace de E^α, stable par P_w, tel que le rayon spectral de $P_{w|F}$ soit strictement inférieur à ρ_w.

Remarques. (a) Si $w(0) = w(1)$, il existe une fonction propre γ de E^α associée à ρ_w, positive ou nulle, telle que $\gamma(0) = \gamma(1)$.

(b) Si la fonction γ du théorème 3.1 est strictement positive, il existe $c > 0$ telle que $1 \leq c\gamma$, d'où $\rho_w^{-n} P_w^n 1 \leq c\gamma$. Il en résulte que $\nu(\rho_w) = 1$.

(c) Si $\nu(\rho_w) = 1$, les valeurs propres de module égal à ρ_w sont également d'indice 1. La décomposition du théorème 3.1 entraîne que $\sup_{n \geq 1} \rho_w^{-n} ||P_w^n 1||_\infty < +\infty$.

(d) S'il existe une fonction P_w-invariante, continue et strictement positive, on a

$$\sup_{n\geq 1} ||P_w^n 1||_\infty < +\infty.$$

On a donc nécessairement $\rho_w = 1$ et $\nu(\rho_w) = 1$. C'est par exemple le cas si w vérifie $w(\frac{x}{2}) + w(\frac{x}{2} + \frac{1}{2}) = 1$, pour tout $x \in [0,1]$, car alors 1 est P_w-invariante.

Cas polynomial. Dans la suite, pour tout entier $\ell \geq 0$, nous désignerons par $\mathcal{T}_\ell$ l'espace engendré par la famille $\{e^{2i\pi kx}, k = -\ell, \ldots, \ell\}$. Supposons que $w \in \mathcal{T}_\ell$, avec $\ell \geq 1$. Tout polynôme trigonométrique $f(x) = \sum_{k\in\mathbf{Z}} a_k e^{2i\pi kx}$ s'écrit de la manière suivante

$$\begin{aligned} f(x) &= \sum_{k\in\mathbf{Z}} a_{2k} e^{2i\pi 2kx} + \sum_{k\in\mathbf{Z}} a_{2k+1} e^{2i\pi(2k+1)x} \\ &= f_0(2x) + e^{2i\pi x} f_1(2x). \end{aligned}$$

En particulier, on a $w(x) = w_0(2x) + e^{2i\pi x} w_1(2x)$, et un calcul évident montre que

$$P_w f(x) = 2[u_0(x) f_0(x) + e^{2i\pi x} u_1(x) f_1(x)].$$

On déduit aisément de cette dernière formule que P_w laisse invariant $\mathcal{T}_{\ell-1}$. L'opérateur $P = P_{w|\mathcal{T}_{\ell-1}}$ s'identifie, par exemple dans la base $\{e^{2i\pi kx}, k = -\ell+1, \ldots, \ell-1\}$, à une matrice carrée d'ordre $2\ell - 1$ que nous expliciterons dans le §5.

4. Caractérisation des filtres d'échelle de classe $\mathcal{C}^\infty$. Dans ce paragraphe, m_0 est une fonction 1-périodique, de classe $\mathcal{C}^\infty$, telle que $m_0(0) = 1$. Rappelons qu'on a noté $\hat{\phi}(\lambda) = \prod_{k\geq 1} m_0(2^{-k}\lambda)$, et que $\hat{\phi}$ est continue sur $\mathbf{R}$. Si $\hat{\phi} \in L^2(\mathbf{R})$, on note ϕ sa transformée de Fourier inverse. Les espaces E et E^α, et la notion d'indice d'une valeur propre, ont été définis dans le paragraphe précédent. Pour $w \in E$ donnée, on note P_w l'opérateur défini par (12). Rappelons que la condition $m_0(\frac{1}{2}) = 0$ est nécessaire pour (P1), (P2). L'objet de cette partie est de simplifier les solutions de (P1) et (P2) données dans le §2.

THÉORÈME 4.1. *Soit m_0 une fonction 1-périodique, de classe $\mathcal{C}^\infty$, telle que $m_0(0) = 1$. En outre, soient $u(x) = |m_0(x)|^2$, ρ_u le rayon spectral de P_u sur E, et $\nu(\rho_u)$ l'indice de ρ_u pour P_u sur E^1.*

Une condition suffisante pour que $\hat{\phi} \in L^2(\mathbf{R})$ est que $m_0(\frac{1}{2}) = 0$, $\rho_u = 1$ et $\nu(\rho_u) = 1$. En outre, une condition nécessaire et suffisante pour que m_0 soit un filtre d'échelle est que m_0 vérifie $m_0(\frac{1}{2}) = 0$, (9), que $\rho_u = 1$ et $\nu(\rho_u) = 1$, et enfin que tout compact invariant pour m_0 contienne 0 ou 1.

Remarques. Rappelons que, si m_0 n'a qu'un nombre fini $q+1$ de zéros, la condition sur les compacts invariants est équivalente à (11).

Si les conditions du théorème 4.1 sont vérifiées, la fonction $\theta_\phi(x) = \sum_{k\in\mathbf{Z}} |\hat{\phi}(x+k)|^2$ est l'unique fonction P_u-invariante, continue et 1-périodique (à un scalaire multiplicatif près), voir le théorème 2.5.

Démonstration du théorème 4.1. D'après la remarque (c) consécutive au théorème 3.1, si $\rho_u = 1$ et $\nu(\rho_u) = 1$, alors $\sup_{n\geq 1} ||P_u^n 1||_\infty < +\infty$. On montre que $\hat{\phi} \in L^2(\mathbf{R})$ grâce au lemme 2.3, appliqué avec $w = u$, $f = 1$, et au lemme de Fatou.

Supposons maintenant que m_0 soit un filtre d'échelle. Alors m_0 s'annule en $\frac{1}{2}$, satisfait (9), et θ_ϕ est une fonction 1-périodique, de classe $\mathcal{C}^\infty$, P_u-invariante, à valeurs strictement positives (voir §2). En vertu de la remarque d) consécutive au théorème 3.1, il vient que $\rho_u = 1$ et $\nu(\rho_u) = 1$. La propriété sur les compacts invariants résulte de la proposition 2.4.

Réciproquement, si $m_0(\frac{1}{2}) = 0$, $\rho_u = 1$,= $\nu(\rho_u) = 1$, alors $\hat{\phi} \in L^2(\mathbf{R})$, et d'après l'hypothèse sur les compacts invariants et la proposition 2.4, θ_ϕ est stictement positive, donc vérifie (8). □

La condition $m_0(\frac{1}{2}) = 0$ assure en général une factorisation de la forme $m_0(x) = ((1+e^{2i\pi x})/2)^r v(x)$, avec $r \in \mathbf{N}^*$ et v régulière telle que $v(\frac{1}{2}) \neq 0$. Sous cette dernière hypothèse, nous nous proposons de simplifier les critères du théorème précédent.

THÉORÈME 4.2. *Soit m_0 une fonction 1-périodique, de classe C^∞, telle que $m_0(0) = 1$. On suppose en outre que m_0 admet un nombre fini $q+1$ de zéros, et que*

$$m_0(x) = \left(\frac{1+e^{2i\pi x}}{2}\right)^r v(x), \quad \forall x \in [0,1],$$

où $r \in \mathbf{N}^*$, et où v est une fonction lipschitzienne, non nulle en $\frac{1}{2}$. Soit en outre, pour tout $n \geq 1$,

$$S_n = \sum_{k=0}^{2^n-1} |v(\frac{k}{2})|^2 \cdots |v(\frac{k}{2^n})|^2.$$

(i) Si m_0 vérifie (9) et (11), alors la suite $(\frac{S_{n+1}}{S_n})_{n\geq 1}$ converge avec une vitesse exponentielle vers un réel noté β_2.

(ii) Une condition nécessaire et suffisante pour que m_0 soit un filtre d'échelle est que m_0 vérifie (9), (11), et que $\beta_2 < 2^{2r}$.

Remarques. Soit $a(x) = |v(x)|^2$. On montre facilement par récurrence que, pour tout $f \in E$,

$$P_a^n f(x) = \sum_{k=0}^{2^n-1} \left|v\left(\frac{x+k}{2}\right)\right|^2 \cdots \left|v\left(\frac{x+k}{2^n}\right)\right|^2 f\left(\frac{x+k}{2^n}\right),$$

d'où $S_n = P_a^n 1(0)$. On peut d'ailleurs utiliser $S_n(x) = P_a^n 1(x)$ à la place de $S_n(0) = S_n$ dans l'assertion i), mais le calcul de S_n a l'avantage de n'utiliser que les valeurs de v sur les points dyadiques.

Si m_0 admet un nombre infini de zéros, l'assertion (ii) reste valable à condition de remplacer (11) par l'hypothèse générale sur les compacts invariants (cf. le théorème 4.1), et β_2 par le rayon spectral ρ_a de P_a sur E (voir [14], [20]). Nous montrerons ci-dessous que, sous les hypothèses du théorème 4.2, $\rho_a = \beta_2$.

Démonstration du théorème 4.2. On note $u(x) = |m_0(x)|^2$, $a(x) = |v(x)|^2$, P_u et P_a les opérateurs associés respectivement à u et a selon (12), et enfin ρ_a le rayon spectral de P_a sur E.

(i) On a vu que $S_n = P_a^n 1(0)$. Nous allons en fait prouver que, pour tout $x \in [0,1]$, la suite $(P_a^{n+1}1(x)/P_a^n 1(x))_{n\geq 1}$ converge avec une vitesse exponentielle vers ρ_a. D'après le théorème 3.1, il existe une fonction γ de E^1, positive ou nulle sur $[0,1]$, telle que $P_a\gamma = \rho_a\gamma$. En vertu de (11), la fonction a ne possède pas de cycle périodique invariant. D'après une des propriétés rappelées dans le §2, on obtient $\gamma > 0$.

Soit T l'opérateur relativisé de P_a, défini sur E par $Tf(x) = \rho_a^{-1}(\gamma(x))^{-1}P_a(\gamma f)(x)$. Notons que $T = P_w$, où $w(x) = \rho_a^{-1}(\gamma(2x))^{-1}\gamma(x)a(x)$ est telle que $w(x)+w(x+\frac{1}{2}) = 1$. Un résultat prouvé dans [7, p. 309] (voir également [22]) nous assure que, pour tout $f \in E$, la suite de fonctions $\{T^n f, n \geq 1\}$ converge dans E (vers une constante). Donc 1 est l'unique valeur propre de module 1 pour T sur E^1. Si f est telle que $P_a f = \lambda f$,

$\lambda \in \mathbf{C}$, alors $Tg = \rho_a^{-1}\lambda g$, où $g = \frac{f}{\gamma}$. Ainsi, ρ_a est l'unique valeur propre de module ρ_a pour P_a sur E^1. Utilisant la quasi-compacité de P_a sur E^1, il vient que $1 = b\gamma + h$, avec $b \in \mathbf{R}$, et où $h \in E^1$ est telle que $(\rho_a^{-n}||P_a^n h||_\infty)_{n\geq 1}$ converge vers 0 avec une vitesse exponentielle. En outre, il existe une constante $c > 0$ telle que $c\gamma \leq 1$, d'où $c\rho_a^n\gamma \leq P_a^n 1$ pour tout $n \geq 1$. Donc $b \neq 0$. On obtient

$$P_a^n 1(x) = \rho_a^n(b\gamma(x) + v_n(x)),$$

où la suite $(||v_n||_\infty)_{n\geq 1}$ converge vers 0 quand $n \to +\infty$ avec une vitesse exponentielle, ce qui prouve la propriété annoncée. En particulier, $\beta_2 = \rho_a$.

(ii) Soit m_0 un filtre d'échelle. Nous devons prouver que $\beta_2 < 2^{2r}$. On a $\rho_u = 1$ et $\nu(\rho_u) = 1$, grâce au théorème précédent. Or, d'après la remarque (a) consécutive au théorème 3.1 (appliquée ici avec $w = a$), la fonction γ ci-dessus est 1-périodique. La fonction $h(x) = (\sin \pi x)^{2r}\gamma(x)$, qui appartient à E^1, vérifie, grâce à la formule classique $\sin 2x = 2 \sin x \cos x$,

$$P_u h(x) = 2^{-2r}|\sin \pi x|^{2r} P_a\gamma(x) = 2^{-2r}\beta_2 h(x).$$

On a donc $2^{-2r}\beta_2 \leq \rho_u = 1$, d'où $\beta_2 \leq 2^{2r}$. En outre, si on avait $\beta_2 = 2^{2r}$, la fonction 1-périodique h serait P_u-invariante avec $h(0) = h(1) = 0$, d'où d'après le théorème 2.5, $h = h(0)\theta_\phi \equiv 0$, ce qui est absurde.

La preuve de la réciproque utilise la formule classique

$$\prod_{k\geq 1} \cos \frac{\lambda}{2^k} = \frac{\sin \lambda}{\lambda},$$

ainsi que le résultat suivant.

LEMME 4.3. *Soit w une fonction de E^α, où $0 < \alpha \leq 1$, 1-périodique, positive ou nulle. Soit en outre ρ_w le rayon spectral de P_w sur E. On a, pour tout réel b tel que $b > \log_2 \rho_w$,*

$$\int_{-\infty}^{+\infty} \frac{1}{1+|\lambda|^b} \prod_{k\geq 1} w(\frac{\lambda}{2^k}) d\lambda < +\infty.$$

Preuve du lemme. Il suffit de vérifier que

$$\sum_{n\geq 0} \int_{2^{n-2}\leq|\lambda|\leq 2^{n-1}} \frac{1}{|\lambda|^b} \prod_{k\geq 1} w(\frac{\lambda}{2^k}) d\lambda < +\infty.$$

Or, on a

$$\begin{aligned} I_n &= \int_{2^{n-2}\leq|\lambda|\leq 2^{n-1}} \frac{1}{|\lambda|^b} \prod_{k\geq 1} w(\frac{\lambda}{2^k}) d\lambda \\ &\leq C 2^{-nb} \int_{2^{n-2}\leq|\lambda|\leq 2^{n-1}} \prod_{k\geq 1} w(\frac{\lambda}{2^k}) d\lambda. \end{aligned}$$

Posons $g(\lambda) = \prod_{k\geq 1} w(\frac{\lambda}{2^k}) = g(\frac{\lambda}{2^n}) \prod_{k=1}^n w(\frac{\lambda}{2^k})$. Comme g est continue, nous avons, pour tout réel λ tel que $2^{n-2} \leq |\lambda| \leq 2^{n-1}$, $g(\lambda) \leq C \prod_{k=1}^n w(\frac{\lambda}{2^k})$. Nous en déduisons

que

$$I_n \leq C2^{-nb} \int_{2^{n-2}\leq|\lambda|\leq 2^{n-1}} \prod_{k=1}^{n} w(\frac{\lambda}{2^k})d\lambda$$
$$\leq C2^{-nb} \int_{-2^{n-1}}^{2^{n-1}} \prod_{k=1}^{n} w(\frac{\lambda}{2^k})d\lambda,$$

d'où, d'après le lemme 2.3,

$$I_n \leq C2^{-nb} \int_0^1 P_w^n 1(\lambda)d\lambda$$
$$\leq C_\epsilon 2^{-nb}(\rho_w + \epsilon)^n,$$

où l'on choisit $\epsilon > 0$ tel que $\rho_w + \epsilon < 2^b$. On en déduit bien la propriété annoncée, et finalement le lemme.

Comme $\beta_2 < 2^{2r}$, le lemme appliqué avec $w = a$ et $b = 2r$ donne

$$\int_{-\infty}^{+\infty} |\hat{\phi}(\lambda)|^2 d\lambda = \int_{-\infty}^{+\infty} \prod_{k\geq 1} \left|\cos \frac{\pi\lambda}{2^k}\right|^{2r} \prod_{k\geq 1} a(\frac{\lambda}{2^k})d\lambda$$
$$= \int_{-\infty}^{+\infty} \frac{|\sin \pi\lambda|^{2r}}{|\pi\lambda|^{2r}} \prod_{k\geq 1} a(\frac{\lambda}{2^k})d\lambda$$
$$\leq C \int_{-\infty}^{+\infty} \frac{1}{1+\lambda^{2r}} \prod_{k\geq 1} a(\frac{\lambda}{2^k})d\lambda$$
$$< +\infty.$$

On a montré que $\hat{\phi} \in L^2(\mathbf{R})$. On déduit de la proposition 2.4 que m_0 est un filtre d'échelle. □

5. Caractérisation des filtres d'échelle polynomiaux. Si m_0 est un polynôme trigonométrique, les deux théorèmes précédents fournissent des conditions nécessaires et suffisantes pour que m_0 soit un filtre d'échelle. En particulier, une condition nécessaire est que $m_0(\frac{1}{2}) = 0$. Cependant, dans le cas polynomial, nous allons voir que les valeurs ρ_u, $\nu(\rho_u)$, et β_2 se calculent à l'aide de matrices.

Plus précisément, considérons un polynôme trigonométrique m_0 possédant $q+1$ zéros, et tel que $m_0(0) = 1$, $m_0(\frac{1}{2}) = 0$. Posons, pour fixer les notations,

$$m_0(x) = \frac{1}{2} \sum_{k=m}^{n} h_k e^{2i\pi kx},$$
$$= \left(\frac{1+e^{2i\pi x}}{2}\right)^r v(x),$$

où $r \in \mathbf{N}^*$ et $v(\frac{1}{2}) \neq 0$. Posant $N = n - m$, nous définissons

$$u(x) = |m_0(x)|^2 = \sum_{k=-N}^{N} u_k e^{2i\pi kx}$$
$$= (\cos \pi x)^{2r} a(x)$$

et

$$a(x) = |v(x)|^2 = \sum_{k=-N+r}^{N-r} a_k e^{2i\pi kx}.$$

Nous savons, d'après le §3, que P_u opère sur l'espace de dimension finie $\mathcal{T}_{N-1}$. En fait, un calcul explicite montre que $P = P_{u|\mathcal{T}_{N-1}}$ admet, dans la base $\{e^{-2i\pi(N-1)x}, \ldots, e^{2i\pi(N-1)x}\}$, la matrice carrée d'ordre $2N-1$ suivante:

$$P_0 = [2u_{2i-j}]_{i,j=-N+1,\ldots,N-1}.$$

De la même façon, P_a laisse invariant $\mathcal{T}_{N-r-1}$, et $P_{a|\mathcal{T}_{N-r-1}}$ admet, dans la base $\{e^{-2i\pi(N-r-1)x}, \ldots, e^{2i\pi(N-r-1)x}\}$, la matrice carrée d'ordre $2N-2r-1$ suivante:

$$P_1 = [2a_{2i-j}]_{i,j=-N+r+1,\ldots,N-r-1}.$$

Rappelons que l'indice d'une valeur propre λ de P_0 est le plus petit entier n tel que $\mathrm{Ker}(P_0 - \lambda Id)^n = \mathrm{Ker}(P_0 - \lambda Id)^{n+1}$.

THÉORÈME 5.1. *Les trois propriétés suivantes sont équivalentes.*

1. m_0 *est un filtre d'échelle.*

2. m_0 *vérifie* (9), (11), *et la plus grande valeur propre positive de* P_0 *est égale à* 1, *son indice valant* 1.

3. m_0 *vérifie* (9), (11), *et la plus grande valeur propre positive de* P_1 *est strictement inférieure à* 2^{2r}.

Si ces dernières conditions sont vérifiées, la fonction d'échelle ϕ *est à support compact inclus dans* $[m, n]$, *et* $\theta_\phi(x) = \sum_{k\in\mathbf{Z}} |\hat{\phi}(x+k)|^2$, *qui appartient à* $\mathcal{T}_N$, *est l'unique fonction* P_u*-invariante continue et 1-périodique (à un scalaire multiplicatif près).*

Démonstration du théorème. L'équivalence des trois conditions résulte des théorèmes 4.1 et 4.2, et des propriétés suivantes.

(a) La plus grande valeur propre positive ρ_0 de P_0 est égale au rayon spectral ρ_u de P_u sur E.

(b) L'indice de ρ_0 pour P_0 est égal à 1 si, et seulement si, tel est le cas pour l'indice $\nu(\rho_u)$ de ρ_u pour P_u sur E^1.

(c) La plus grande valeur propre positive de P_1 est égale au rayon spectral $\rho_a = \beta_2$ de P_a sur E.

Soit $\rho(P_0)$ le rayon spectral de la matrice P_0 (ou encore celui de $P_{u|\mathcal{T}_{N-1}}$). Comme la fonction 1 appartient à $\mathcal{T}_{N-1}$, on déduit de (13) que $\rho(P_0) = \rho_u$. En outre, P_u étant un opérateur positif sur $\mathcal{T}_{N-1}$, on sait, d'après la théorie classique des opérateurs positifs en dimension finie [1], que $\rho(P_0)$ est la plus grande valeur propre positive de $P_{u|\mathcal{T}_{N-1}}$, donc de la matrice P_0. La propriété (a) est prouvée, et le (c) s'établit de la même façon.

Le (b) découle des remarques consécutives au théorème 3.1, et des résultats classiques sur les opérateurs positifs en dimension finie : pour simplifier, supposons que $\rho_u = 1$, et par conséquent que $\rho_0 = 1$. Si $\nu(\rho_u) = 1$, alors $\sup_{n\geq 1} ||P_u^n 1||_\infty = \sup_{n\geq 1} ||(P_{u|\mathcal{T}_{N-1}})^n(1)||_\infty < +\infty$. Donc l'indice de ρ_0 pour P_0 est égal à 1.

Réciproquement, si cette dernière propriété est satisfaite, on sait que toutes les valeurs propres de module 1 pour $P_{u|\mathcal{T}_{N-1}}$, donc pour P_0, sont d'indice 1. D'où

$$\sup_{n\geq 1} ||(P_{u|\mathcal{T}_{N-1}})^n(1)||_\infty = \sup_{n\geq 1} ||P_u^n 1||_\infty < +\infty,$$

et finalement $\nu(\rho_u) = 1$.

Le fait que ϕ soit à support compact inclus dans $[m, n]$ est une propriété classique des filtres d'échelle polynomiaux démontrée dans [25], [8]. On a $\theta_\phi \in \mathcal{T}_N$ d'après la formule sommatoire de Poisson. Enfin le résultat d'unicité résulte du théorème 2.5. □

Remarque. Si les coefficients de fourier u_k de u sont réels, les opérateurs P_u et P_a opèrent respectivement sur les espaces $\tilde{\mathcal{T}}_{N-1} = \text{vect}\{1, \ldots, \cos 2\pi(N-1)x\}$ et $\tilde{\mathcal{T}}_{N-r-1} = \text{vect}\{1, \ldots, \cos 2\pi(N-r-1)x\}$. Le théorème ci-dessus reste alors valable quand on remplace P_0 et P_1 respectivement par $Q_0 = P_{u|\tilde{\mathcal{T}}_{N-1}}$ et $Q_1 = P_{a|\tilde{\mathcal{T}}_{N-r-1}}$.

6. Résolution de (P3) et exemples. Les différentes méthodes (directes ou fréquentielles) et les travaux antérieurs, relatifs à la régularité de ϕ, ont été indiqués dans le §1.

Soit un réel $b > 0$. Une fonction f est de classe $\mathcal{C}^b$ sur $\mathbf{R}$ si f est $[b]$-fois continuement dérivable, et s'il existe une constante $c > 0$ telle que l'on ait, pour tout couple (x, y) de réels,

$$|f^{[b]}(y) - f^{[b]}(x)| \leq c|y - x|^{b-[b]}.$$

Soit p un réel tel que $p \geq 1$. Si $\hat{\phi} \in L^p(\mathbf{R})$, nous posons

$$s_p = \sup\{s > 0 : \int_{-\infty}^{+\infty} |\hat{\phi}(\lambda)|^p (1 + |\lambda|^{ps}) d\lambda < +\infty\}.$$

Rappelons les inclusions classiques entre les espaces de Sobolev et les espaces d'Hölder : si $s_2 > \frac{1}{2}$, ϕ est de classe $\mathcal{C}^{s_2 - \frac{1}{2} - \epsilon}$ pour tout $\epsilon > 0$. En général s_2 n'est pas le coefficient d'hölder optimal α_0 de ϕ. En revanche α_0 est inférieur à $s_2 + \frac{1}{2}$. Par ailleurs, si ϕ est intégrable et si $s_1 > 0$, alors ϕ est de classe $\mathcal{C}^{s_1 - \epsilon}$ pour tout $\epsilon > 0$.

Nous nous proposons ici de calculer s_1, s_2, et plus généralement s_p, dans le cas général où m_0 est de classe $\mathcal{C}^\infty$. Mais commençons par donner une première condition pour que $\hat{\phi} \in L^p(\mathbf{R})$, laquelle sera simplifiée dans le théorème 6.2.

THÉORÈME 6.1. *Soit m_0 une fonction 1-périodique, de classe $\mathcal{C}^\infty$, telle que $m_0(0) = 1$. Soient en outre p un réel tel que $p \geq 1$, $w = |m_0|^p$, ρ_p le rayon spectral de P_w, et enfin $\nu(\rho_p)$ l'indice de ρ_p pour P_w sur E^{α_p}, où α_p est la partie fractionnaire de p si p n'est pas entier, et $\alpha_p = 1$ sinon. Alors une condition suffisante pour que $\hat{\phi} \in L^p(\mathbf{R})$ est que $m_0(\frac{1}{2}) = 0$, $\rho_p = 1$ et $\nu(\rho_p) = 1$. C'est une condition nécessaire et suffisante sous l'hypothèse (9) et si tout compact invariant pour m_0 contient 0 ou 1.*

Démonstration. La première propriété s'établit comme dans le théorème 4.1 avec w et E^{α_p} à la place de $|m_0|^2$ et E^1.

Réciproquement, supposons que $\hat{\phi} \in L^p(\mathbf{R})$, et que tout compact invariant pour m_0 contient 0 ou 1. Soit $\theta_p(x) = \sum_{k \in \mathbf{Z}} |\hat{\phi}(x + k)|^p$. La fonction θ_p est semi-continue inférieurement, P_w-invariante (utiliser (3) et considérer ici P_w sur $L^1([0, 1])$, et strictement positive d'après la remarque consécutive à la proposition 2.4. Si en outre θ_p est continue, ou plus généralement bornée, on a nécessairement $\rho_u = 1$ et $\nu(\rho_u) = 1$, voir remarque (d) consécutive au théorème 3.1. Cependant, dans le cas général, la continuité de θ_p n'est pas assurée a priori. On peut alors procéder de la manière suivante: considérons la fonction γ de E^{α_p}, positive ou nulle, vérifiant $P_w\gamma = \rho_p\gamma$, dont l'existence est assurée par le théorème 3.1. Comme $m_0(\frac{1}{2}) = 0$, on a $P_w^n 1(0) = 1$, pour tout entier $n \geq 1$, et donc $\rho_p \geq 1$ en vertu de (13). Par ailleurs, il existe une constante $c > 0$ telle que $\gamma \leq c\theta_p$, d'où, pour tout entier $n \geq 1$, $\rho_p^n \gamma \leq \theta_p$. Comme θ_p est intégrable sur $[0, 1]$, on a nécessairement $\rho_p = 1$.

Pour prouver que $\nu(\rho_p) = 1$, procédons par l'absurde : si $\nu(\rho_p) \geq 2$, il existe deux fonctions f et g de E^{α_p}, non identiquement nulles, telles que $P_w f = f + g$ et $P_w g = g$. Là encore, nous avons $|f| \leq c\theta_p$, où $c > 0$, d'où, pour tout entier $n \geq 1$, $|P_w^n f| = |f + ng| \leq P_w^n |f| \leq c\theta_p$. Du fait que $\theta_p \in L^1([0,1])$, la contradiction s'établit là encore en faisant tendre n vers $+\infty$. □

L'objet du théorème suivant est de simplifier les conditions précédentes, et de calculer les coefficients s_p définis au début de ce paragraphe. A cet effet, on supposera que le nombre de zéros de m_0 est fini, un énoncé plus général étant donné dans [20].

THÉORÈME 6.2. *Soit m_0 une fonction 1-périodique, de classe C^∞, admettant un nombre fini $q+1$ de zéros, et vérifiant $m_0(0) = 1$, (9) et (11). On suppose en outre que*

$$m_0(x) = \left(\frac{1 + e^{2i\pi x}}{2}\right)^r v(x), \quad \forall x \in [0,1],$$

où $r \in \mathbf{N}^*$, et où v est une fonction lipschitzienne, non nulle en $\frac{1}{2}$. Soit en outre, pour p quelconque tel que $p \geq 1$, et tout $n \geq 1$,

$$S_{p,n} = \sum_{k=0}^{2^n-1} |v(\frac{k}{2})|^p \dots |v(\frac{k}{2^n})|^p.$$

(i) La suite $(S_{p,n+1}/S_{p,n})_{n\geq 1}$ converge avec une vitesse exponentielle vers un réel β_p.

(ii) Une condition nécessaire et suffisante pour que $\hat{\phi} \in L^p(\mathbf{R})$ est que $\beta_p < 2^{pr}$. On a alors

$$s_p = r - \frac{1}{p} \log_2 \beta_p.$$

On notera le parallèle avec l'énoncé du théorème 4.2. En particulier, $S_n = S_{2,n}$, et

$$P^n_{|v|^p} f(x) = \sum_{k=0}^{2^n-1} |v(\frac{x+k}{2})|^p \dots |v(\frac{x+k}{2^n})|^p f(\frac{x+k}{2^n}),$$

d'où $S_{p,n} = P^n_{|v|^p} 1(0)$. La démonstration du (i) est exactement la même que dans le théorème 4.2. En revanche, la démonstration du (ii) étant un peu plus technique, nous l'avons reportée en appendice.

Remarques. Soit m_0 un filtre d'échelle de classe C^∞. On sait que ϕ est intégrable sur $\mathbf{R}$ [25]. Par conséquent si $s_1 > 0$, ϕ est de classe $C^{s_1-\epsilon}$ pour tout $\epsilon > 0$. Dans le cas où m_0 est un polynôme trigonométrique positif ou nul, on montre dans [5] que, si $\hat{\phi}$ ne s'annule pas sur $[-\frac{1}{2}, +\frac{1}{2}]$, alors s_1 est le coefficient de Hölder optimal pour ϕ (si $s_1 \in \mathbf{N}^*$, il faut ici remplacer la classe de Hölder par la classe de Zygmund).

Cas particulier où m_0 est un polynôme trigonométrique. Nous conservons les notations du §5. Soit p un entier positif quelconque.

Si p est pair, $|v|^p$ est un polynôme trigonométrique, $P_{|v|^p}$ laisse invariant l'espace de dimension finie $\mathcal{T}_{\frac{p}{2}(N-r)-1}$, et β_p est aussi la plus grande valeur propre positive de l'opérateur $P_{|v|^p}$ restreint à l'espace $\mathcal{T}_{\frac{p}{2}(N-r)-1}$.

Si v est positif ou nul, cette dernière remarque s'étend au cas où p est impair. On dispose alors d'un critère très simple pour que $\hat{\phi} \in L^1(\mathbf{R})$.

Exemples. Nous donnons les valeurs de s_1 et s_2 pour les filtres (de longueur infinie) présentés dans [16]:

• (filtre de Butterworth). $m_0(x) = (\frac{1+e^{2i\pi x}}{2})^2 \left|8(6+2\cos 4\pi x)^{-1}\right|^{\frac{1}{2}}$.
$s_1 \simeq 0.7633$ et $s_2 \simeq 1.2564$.

• (filtre de Butterworth). $m_0(x) = (\frac{1+e^{2i\pi x}}{2})^3 \left|32(20+12\cos 4\pi x)^{-1}\right|^{\frac{1}{2}}$.
$s_1 \simeq 1.5615$ et $s_2 \simeq 2.045$.

• $m_0(x) = (\frac{1+e^{2i\pi x}}{2})^4 \left|256(4-\cos 2\pi x)(448+320\cos 4\pi x)^{-1}\right|^{\frac{1}{2}}$.
$s_1 \simeq 2.238$ et $s_2 \simeq 2.702$.

• (filtre de Butterworth). $m_0(x) = (\frac{1+e^{2i\pi x}}{2})^4 \left|128(70+56\cos 4\pi x+2\cos 8\pi x)^{-1}\right|^{\frac{1}{2}}$.
$s_1 \simeq 2.3707$ et $s_2 \simeq 2.843$.

• $m_0(x) = (\frac{1+e^{2i\pi x}}{2})^5 \left|512(34-20\cos 2\pi x+2\cos 4\pi x)(4608+3584\cos 4\pi x)^{-1}\right|^{\frac{1}{2}}$.
$s_1 \simeq 2.843$ et $s_2 \simeq 3.282$.

• (filtre de Butterworth). $m_0(x) = (\frac{1+e^{2i\pi x}}{2})^5 \left|512(252+240\cos 4\pi x+20\cos 8\pi x)^{-1}\right|^{\frac{1}{2}}$
$s_1 \simeq 3.184$ et $s_2 \simeq 3.649$.

7. Application à la construction d'interpolations dyadiques continues. Soit $(G_s)_{s\in\mathbf{N}}$ une famille croissante de sous-groupes discrets de $\mathbf{R^n}$ tels que $G_\infty = \cup_{s\geq 0} G_s$ soit dense dans $\mathbf{R^n}$. Soit en outre f une fonction quelconque, à valeurs complexes, définie sur G_0. L'interpolation [2], [10], est un schéma qui permet de prolonger f, de manière itérative, à $G_1, G_2, \ldots, G_n, \ldots$: on obtient ainsi une fonction f définie sur G_∞. L'une des questions est de caractériser les interpolations, dites continues, telles que toute fonction interpolée sur G_∞ admette un prolongement continu sur $\mathbf{R^n}$ (voir [11]).

Nous travaillons ici avec $n = 1$ et dans le cadre dyadique $G_s = 2^{-s}\mathbf{Z}$, $\mathbf{s} \in \mathbf{N}$. Après quelques rappels, nous nous proposons de donner, sous des hypothèses assez générales, un critère très simple pour qu'une interpolation dyadique soit continue.

Procédé d'interpolation dyadique. Nous désignerons par D l'ensemble des réels dyadiques, c'est-à-dire de la forme $2^{-r}k$, où $r \in \mathbf{N}$ et $k \in \mathbf{Z}$.

Etant donné une famille $\{c(n),\ n \in \mathbf{Z}\}$ de réels, nuls sauf pour un nombre fini d'entiers n, on définit le procédé $(\mathcal{D})$ d'interpolation dyadique de la manière suivante: à toute suite réelle $a = (a(n))_{n\in\mathbf{Z}}$, on associe la fonction f, définie sur D par

$f(n) = a(n)$ si $n \in \mathbf{Z}$

$f(2^{-r}n + 2^{-(r+1)}) = \sum_{k\in\mathbf{Z}} c(n-k) f(2^{-r}k)$ pour $r = 0, 1, 2, \ldots$, et $n \in \mathbf{Z}$.

Pour fixer les idées, nous noterons dans la suite p et q, $p \leq q$, les deux entiers tels que l'on ait $c(n) = 0$ pour $n \notin [p, q]$ et $c(p) \neq 0, c(q) \neq 0$.

Notons D_r, pour $r \in \mathbf{N}$, l'ensemble $\{2^{-r}k,\ k \in \mathbf{Z}\}$ des dyadiques d'ordre r. Les éléments de $D_{r+1} - D_r$ sont exactement les réels de la forme $2^{-r}n + 2^{-(r+1)}$, $n \in \mathbf{Z}$. Ainsi, le procédé $(\mathcal{D})$ est bien défini dans le sens qu'il permet, partant d'une suite a quelconque, de construire une et une seule fonction f définie sur D. On dit que f est la *fonction interpolée par $(\mathcal{D})$ partant de a.*

Remarque. Il est naturel d'imposer que la fonction interpolée partant de la suite $\{a(n) = 1,\ n \in \mathbf{Z}\}$ soit identiquement égale à 1 sur D, c'est-à-dire que

$$\sum_{n\in\mathbf{Z}} c(n) = 1.$$

Interpolante fondamentale. Soit $(\mathcal{D})$ un procédé d'interpolation dyadique. La fonction ϕ_c définie sur D comme l'interpolée par $(\mathcal{D})$ partant de la suite $\delta_0 = (\delta_{0,n})_{n\in\mathbf{Z}}$

est appelée *l'interpolante fondamentale* relative à $(\mathcal{D})$ (on a noté $\delta_{0,n} = 1$ si $n = 0$, $\delta_{0,n} = 0$ sinon).

Reprenant les notations ci-dessus, l'interpolante fondamentale ϕ_c est à support borné contenu dans $[2p+1, 2q+1]$, autrement dit on a $\phi_c(x) = 0$ pour tout dyadique x n'appartenant pas à ce dernier intervalle. En outre, toute fonction interpolée par $(\mathcal{D})$ s'écrit sous la forme

$$f(x) = \sum_{k \in \mathbf{Z}} f(k)\phi_c(x-k), \quad x \in D.$$

Pour tout dyadique x fixé, la série ci-dessus est une somme finie. En particulier, si a est une suite à support fini, la fonction interpolée correspondante est à support borné.

Interpolation dyadique continue. Soit $(\mathcal{D})$ un procédé d'interpolation dyadique. Puisque D est dense dans $\mathbf{R}$, la question du prolongement à $\mathbf{R}$ des interpolées a un sens, ce qui conduit à la définition suivante: On dit que $(\mathcal{D})$ est une *interpolation dyadique continue* si l'interpolante fondamentale ϕ_c relative à $(\mathcal{D})$ admet un prolongement continu sur $\mathbf{R}$, que l'on notera encore ϕ_c.

D'après ce qui précède, ϕ_c est alors à support compact contenu dans $[2p+1, 2q+1]$, et toute fonction f interpolée par $(\mathcal{D})$ admet un prolongement $\tilde{f}$ continu sur $\mathbf{R}$, donné par

$$\tilde{f}(x) = \sum_{k \in \mathbf{Z}} f(k)\phi_c(x-k) \quad \text{pour tout } x \in \mathbf{R}. \tag{14}$$

Si l'interpolante fondamentale ϕ_c est de classe $\mathcal{C}^p$ sur $\mathbf{R}$, toute fonction f interpolée admet un prolongement de classe $\mathcal{C}^p$, les dérivées successives jusqu'à l'ordre p de f s'obtenant par dérivation terme à terme dans (14).

Lien avec les filtres d'échelle et applications. Dans ce travail, nous nous limiterons aux interpolations dyadiques continues telles que l'interpolante fondamentale engendre par translations entières un système de Riesz (§1). Dans ce cas, si $(f(n))_{n \in \mathbf{Z}}$ est une suite de $\ell^2(\mathbf{Z})$, la série dans la formule (14) converge, non seulement ponctuellement, mais également dans $L^2(\mathbf{R})$. Nous verrons qu'en fait la plupart des interpolations dyadiques continues satisfont à l'hypothèse précédente.

THÉORÈME 7.1. *Soit $\{c(p), \ldots, c(q)\}$ une famille de réels tels que $\sum_k c(k) = 1$, et soit $(\mathcal{D})$ le procédé d'interpolation dyadique associé. Les deux propriétés suivantes sont équivalentes:*

1. *$(\mathcal{D})$ est une interpolation dyadique continue, et son interpolante fondamentale ϕ_c engendre par translations entières un système de Riesz.*
2. *Le polynôme trigonométrique*

$$H_c(x) = \frac{1}{2} + \frac{1}{2}\sum_{k=p}^{q} c(k)e^{2i\pi(2k+1)x}$$

est un filtre d'échelle, et la fonction d'échelle ϕ associée est continue, telle que

$$\phi(n) = \delta_{0,n}, \quad \forall n \in \mathbf{Z}. \tag{15}$$

On a alors $\phi_c = \phi$.

Démonstration du théorème. Pour simplifier les notations, nous posons $h_{2n} = \delta_{0,n}$, et $h_{2n+1} = c(n)$, $n \in \mathbf{Z}$.

Pour prouver que 1 implique 2, il suffit de vérifier que l'interpolante fondamentale ϕ_c satisfait à l'équation d'échelle (voir §1)

$$\phi_c(\frac{x}{2}) = \sum_{n\in\mathbf{Z}} h_n \phi_c(x-n), \qquad x \in \mathbf{R}. \tag{16}$$

Par continuité, il suffit de prouver que $f : x \to \phi_c(\frac{x}{2})$ et $g : x \to \sum_{n\in\mathbf{Z}} h_n \phi_c(x-n)$ coïncident sur D. Or $\phi_c(\cdot - n)$ est la fonction interpolée par $(\mathcal{D})$ partant de la suite δ_n, définie par $\delta_n(k) = \delta_{0,n-k}$. Par linéarité, la fonction g est donc la fonction interpolée par $(\mathcal{D})$ partant de la suite $a = \sum_{n\in\mathbf{Z}} h_n \delta_n$. D'autre part, pour tout entier n et tout entier $r \geq 0$, nous obtenons, d'après la définition du procédé d'interpolation,

$$\begin{aligned} f(2^{-r}n + 2^{-(r+1)}) &= \phi_c(2^{-(r+1)}n + 2^{-(r+2)}) \\ &= \sum_{k\in\mathbf{Z}} c(n-k)\phi_c(2^{-(r+1)}k) \\ &= \sum_{k\in\mathbf{Z}} c(n-k) f(2^{-r}k). \end{aligned}$$

Donc f est aussi une fonction interpolée par $(\mathcal{D})$. Pour prouver que f et g coïncident sur D, il suffit de montrer que $f = g$ sur $\mathbf{Z}$. Or, on a

$$\begin{aligned} g(k) &= h_k &&= \begin{cases} \delta_{0,p} & \text{si } k = 2p, \\ c(p) & \text{si } k = 2p+1, \end{cases} \\ f(k) &= \phi_c(\tfrac{k}{2}) &&= \begin{cases} \delta_{0,p} & \text{si } k = 2p, \\ f(2p+1) = \phi_c(p+\frac{1}{2}) = c(p) & \text{si } k = 2p+1. \end{cases} \end{aligned}$$

Démontrons maintenant que 2 entraîne 1. Les fonctions $x \to \phi(2^{-r}x)$, $r \in \mathbf{N}$, qui appartiennent à l'espace engendré par le système de Riesz $\{\phi(\cdot - k),\ k \in \mathbf{Z}\}$ (voir § 1), s'écrivent, grâce à (15), sous la forme

$$\phi(2^{-r}x) = \sum_{k\in\mathbf{Z}} \phi(2^{-r}k)\phi(x-k),$$

d'où, pour $x = n + \frac{1}{2}$, $n \in \mathbf{Z}$,

$$\phi(2^{-r}n + 2^{-(r+1)}) = \sum_{k\in\mathbf{Z}} \phi\left(\frac{2(n-k)+1}{2}\right)\phi(2^{-r}k).$$

Or, grâce à l'équation d'échelle (4) et à (15), nous obtenons pour tout entier p,

$$\phi(\frac{2p+1}{2}) = \sum_{k\in\mathbf{Z}} h_n \phi(2p+1-n) = h(2p+1) = c(p).$$

Ainsi, ϕ apparaît comme l'interpolante fondamentale associée à la famille $\{c(n),\ n \in \mathbf{Z}\}$, ce qui prouve la réciproque. □

Remarque. Supposons que $\{c(p), \ldots, c(q)\}$ définisse une interpolation dyadique continue, et notons ϕ_c son interpolante fondamentale. Nous avons montré ci-dessus, sans aucune hypothèse supplémentaire, que $\hat{\phi}_c$ vérifie l'équation $\hat{\phi}_c(\lambda) = H_c(\frac{\lambda}{2})\hat{\phi}_c(\frac{\lambda}{2})$ (utiliser la transformée de Fourier dans (16)). Par conséquent, en vertu des résultats précédents, ϕ_c engendre par translations entières un système de Riesz si, et seulement si, H_c ne possède pas de cycle périodique autre que $\{0\}$ et $\{1\}$, ce qui est presque

toujours vérifié. Ceci prouve bien que les interpolations dyadiques continues étudiées dans ce paragraphe sont assez générales.

Nous nous proposons maintenant de caractériser, par une condition très simple, les interpolations dyadiques du théorème précédent vérifiant en outre la propriété suivante: $\hat{\phi}_c \in L^1(\mathbf{R})$. Si l'on souhaite construire des interpolations dyadiques telles que ϕ_c soit assez régulière, cette dernière condition est naturelle.

Soient $\{c(p), \dots, c(q)\}$ une famille de réels tels que $\sum_k c(k) = 1$, $(\mathcal{D})$ le procédé d'interpolation dyadique associé, et soit H_c le polynôme trigonométrique défini dans le théorème précédent. Comme $H_c(\frac{1}{2}) = 0$, il existe $r \in \mathbf{N}^*$ et un polynôme trigonométrique v, avec $v(\frac{1}{2}) \neq 0$, tels que

$$H_c(x) = \left(\frac{1 + e^{2i\pi x}}{2}\right)^r v(x).$$

THÉORÈME 7.2. *On suppose que la fonction H_c possède $q+1$ zéros, et qu'elle satisfait* (11), *et $|H_c(x)| + |H_c(x + \frac{1}{2})| > 0$ pour tout $x \in \mathbf{R}$. Soit, pour $n \geq 1$,*

$$S_n = \sum_{k=0}^{2^n-1} \left|v\left(\frac{k}{2}\right)\right| \dots \left|v\left(\frac{k}{2^n}\right)\right|.$$

La suite $(S_{n+1}/S_n)_{n\geq 1}$ converge avec une vitesse exponentielle vers un réel β. Les deux propriétés suivantes sont équivalentes.

1. *$(\mathcal{D})$ est une interpolation dyadique continue, $\hat{\phi}_c \in L^1(\mathbf{R})$, et ϕ_c engendre par translations entières un système de Riesz.*
2. *$\beta < 2^r$.*

La fonction ϕ_c est alors de classe $\mathcal{C}^{r-\log_2 \beta-\epsilon}$ pour tout $\epsilon > 0$.

Remarques. La démonstration du théorème 7.2 utilise les résultats du §6. En particulier, d'après le théorème 6.2, le coefficient s_1, relatif à ϕ_c, est donné par $s_1 = r - \log_2 \beta$, ce qui prouve l'estimation ci-dessus sur la régularité de ϕ_c.

Soient $w = |H_c|$, P_w l'opérateur défini par (12), ρ_w le rayon spectral de P_w sur E, et enfin $\nu(\rho_w)$ l'indice de ρ_w pour P_w sur E^1 (voir §3). Sous les hypothèses du théorème 7.2, la condition $\beta < 2^r$ est équivalente à $\rho_w = 1$ et $\nu(\rho_w) = 1$ (voir §6).

Notons que $H_c(\frac{x}{2}) + H_c(\frac{x}{2} + \frac{1}{2}) = 1$ pour tout $x \in [0, 1]$. D'après la remarque (d) consécutive au théorème 3.1, si H_c est positive ou nul, on a $\rho_w = 1$ et $\nu(\rho_w) = 1$. Le résultat de [11], selon lequel toute famille $\{c(p), \dots, c(q)\}$ telle que $H_c \geq 0$ est une interpolation dyadique continue, est une simple conséquence de la remarque précédente.

Si v est positif, il est de la forme

$$v(x) = \sum_{k=-\ell-1}^{\ell+1} a_k e^{2i\pi kx},$$

où $\ell \in \mathbf{N}^*$. Le nombre β du théorème 7.2 est alors égal à la plus grande valeur propre positive de la restriction de P_v à l'espace de dimension finie $\mathcal{T}_\ell$, voir §§3 et 5.

La définition des interpolations dyadiques conserve un sens lorsque la suite $\{c(n), n \in \mathbf{Z}\}$ n'est pas à support fini (mais suffisamment décroissante quand $|n| \to +\infty$). Les résultats du paragraphe 6 étant énoncés pour les filtres de classe $\mathcal{C}^\infty$, le théorème précédent reste donc valable si $H_c \in \mathcal{C}^\infty$.

On peut généraliser la notion d'interpolation dyadique de la manière suivante (voir les notations du début du paragraphe) : soit $m \in \mathbf{N}^*$. Partant de fonctions a_α, définies sur G_0, indexées par les multi-indices $\alpha, |\alpha| \leq m$, on veut construire itérativement une fonction f sur G_∞, qui admette un prolongement $\tilde{f}$ de classe $\mathcal{C}^m$ sur $\mathbf{R^n}$, et tel que les fonctions $\frac{\partial^\alpha \tilde{f}}{\partial x^\alpha}$ coïncident avec a_α sur G_0, voir [24]. Dans le cadre $n = 1$ et $G_s = 2^{-s}\mathbf{Z}$, la description pour $m \geq 1$ est formellement identique au cas $m = 0$ étudié ici, et conduit aux analyses multirésolutions de multiplicité $m + 1$ [19].

Démonstration du théorème 7.2. On pose

$$\hat{\phi}(\lambda) = \prod_{k=1}^{+\infty} H_c\left(\frac{\lambda}{2^k}\right).$$

D'après le théorème 6.2, les conditions $\beta < 2^r$ et $\hat{\phi} \in L^1(\mathbf{R})$ sont équivalentes. Or si le point 1 du théorème est vérifié, on a en particulier $\hat{\phi}_c \in L^1(\mathbf{R})$, et l'on sait, grâce au théorème précédent, que $\hat{\phi}_c = \hat{\phi}$, donc $\beta < 2^r$.

Réciproquement, supposons que $\beta < 2^r$. Alors $\hat{\phi} \in L^1(\mathbf{R})$. La transformée de Fourier inverse ϕ de $\hat{\phi}$ est donc continue, à support compact, et d'après (11), les translatées entières de ϕ forme un système de Riesz. Par conséquent, H_c est un filtre d'échelle. En vertu du théorème 7.1, il reste à prouver que ϕ vérifie (15), ou ce qui revient au même, que le polynôme trigonométrique $\gamma(x) = \sum_{n\in\mathbf{Z}} \phi(n)e^{2i\pi nx}$ est identiquement égal à 1.

Grâce à la formule sommatoire de Poisson, on a $\gamma(x) = \sum_{k\in\mathbf{Z}} \hat{\phi}(x + k)$. On a donc $\gamma(x) = \sum_{k\in\mathbf{Z}} H_c(\frac{x}{2} + \frac{k}{2})\hat{\phi}(\frac{x}{2} + \frac{k}{2})$. Découpant cette dernière somme suivant les indices k pairs et impairs, nous obtenons la formule

$$\gamma(x) = H_c\left(\frac{x}{2}\right)\gamma\left(\frac{x}{2}\right) + H_c\left(\frac{x}{2} + \frac{1}{2}\right)\gamma\left(\frac{x}{2} + \frac{1}{2}\right).$$

D'autre part nous avons $H_c(\frac{x}{2}) + H_c(\frac{x}{2} + \frac{1}{2}) = 1$ pour tout $x \in [0, 1]$. Ceci nous conduit à considérer l'opérateur P_{H_c} défini selon (12) (avec H_c à la place de w) sur l'espace E des fonctions continues sur $[0, 1]$.

Les fonctions 1 et γ sont P_{H_c}-invariantes. Nous allons conclure par l'absurde : supposons que γ ne soit pas identiquement égale à 1. On peut alors choisir un réel a tel que la fonction $f = \gamma + a1$, qui reste invariante par P_{H_c}, soit nulle en 0. Cette dernière est donc de la forme

$$f(x) = e^{i\pi\ell x} \sin^\ell \pi x \; g(x),$$

où $\ell \in \mathbf{N}^*$ et $g(0) \neq 0$. Explicitant l'équation $P_{H_c} f = f$, nous obtenons l'égalité

$$\begin{aligned} e^{i\pi\ell x} \sin^\ell(\pi x) g(x) = {} & e^{i(\ell+r)\frac{\pi}{2}x} g\left(\frac{x}{2}\right) v\left(\frac{x}{2}\right) \sin^\ell\left(\frac{\pi}{2}x\right) \cos^r\left(\frac{\pi}{2}x\right) \\ & + e^{i(\ell+r)\pi(\frac{x}{2}+\frac{1}{2})} g\left(\frac{x}{2} + \frac{1}{2}\right) v\left(\frac{x}{2} + \frac{1}{2}\right) \sin^r\left(\frac{\pi}{2}x\right) \cos^\ell\left(\frac{\pi}{2}\right). \end{aligned}$$

Utilisant la relation $2 \sin x \cos x = \sin 2x$, on montre aisément que nécessairement $\ell \geq r$. La fonction f s'écrit finalement sous la forme

$$f(x) = e^{i\pi rx} \sin^r \pi x f_0(x),$$

et la relation $P_{H_c} f = f$ devient

$$\begin{aligned} f(x) &= e^{i\frac{\pi}{2}rx} \sin^r \frac{\pi}{2}x \; f_0\left(\frac{x}{2}\right) \cos^r \frac{\pi}{2}x \; v\left(\frac{x}{2}\right) \\ &\quad + (-1)^r e^{i\pi(x+1)r} \cos^r \frac{\pi}{2}x \; f_0\left(\frac{x}{2}\right) \sin^r \frac{\pi}{2}x \; v\left(\frac{x}{2}+\frac{1}{2}\right) \\ &= 2^{-r} e^{i\pi rx} \sin^r \pi x \; P_v f_0(x), \end{aligned}$$

où P_v est l'opérateur défini selon (12). On a prouvé que $P_v f_0 = 2^r f_0$. Or pour tout $f \in E$, pour tout $x \in [0,1]$, et tout entier $n \geq 1$, on a $|P_v^n f(x)| \leq P_{|v|}^n |f|(x)$, où $P_{|v|}$ est l'opérateur associé à $|v|$ selon (12). On a démontré dans le §6 que β est le rayon spectral de $P_{|v|}$ sur E. On en déduit que le rayon spectral ρ_v de P_v sur E est tel que $\rho_v \leq \beta < 2^r$. Par conséquent 2^r ne peut être une valeur propre de P_v sur E. □

8. Appendice: Démonstration du théorème 6.2. Il suffit de traiter le cas $p = 1$. Pour p quelconque, on remplacera $|m_0|$ par $|m_0|^p$, et E^1 par E^{α_p} (α_p a été défini dans le théorème 6.1). On pose $u = |m_0|$, $a = |v|$. On note P_u, P_a les opérateurs définis sur E selon (12), et ρ_u, ρ_a leur rayon spectral. Soit enfin $\nu(\rho_u)$ l'indice de ρ_u pour P_u sur E^1. La propriété (i) s'établit comme dans le théorème 4.2. En particulier, on a $\rho_a = \beta_1$. Pour prouver (ii), nous aurons besoin du lemme suivant.

LEMME 8.1. *Sous les hypothèses du théorème* 6.2, *si* $\rho_u = 1$ *et* $\nu(\rho_u) = 1$, *alors il existe une fonction* $\gamma \in E^1$ P_u*-invariante,* 1*-périodique, à valeurs strictement positives, et* γ *est l'unique fonction* P_u*-invariante,* 1*-périodique et continue (à un scalaire multiplicatif près).*

Démonstration du lemme. Commençons par prouver qu'il existe une fonction $\gamma \in E^1$, P_u-invariante, 1-périodique, à valeurs positives ou nulles, telle que $\gamma(0) = 1$. Comme $m_0(\frac{1}{2}) = 0$, on a $P_u^n 1(0) = 1$ pour tout entier $n \geq 1$. D'autre part, d'après le théorème 3.1, on a $M = \sup_{n\geq 1} ||P_u^n 1||_\infty < +\infty$, et on montre aisément par récurrence (voir [18]) que, pour tout entier $n \geq 1$ et toute fonction $f \in E^1$,

$$|||P_u^n f|||_1 \leq 2^{-n} ||P_u^n 1||_\infty |||f|||_1 + R_n ||f||_\infty, \tag{17}$$

où R_n est une constante positive ne dépendant que de n et u (la norme $|||\cdot|||_1$ a été définie dans le § 3). On en déduit qu'il existe un entier N pour lequel on a, pour tout $f \in E^1$,

$$|||P_u^N f|||_1 \leq r|||f|||_1 + R_N ||f||_\infty, \quad \text{où} \quad 0 < r < 1.$$

D'où,

$$\begin{aligned} |||P_u^{2N} f|||_1 &= |||P_u^N (P_u^N f)|||_1 \\ &\leq r|||P_u^N f|||_1 + R_N ||P_u^N f||_\infty \\ &\leq r^2 |||f|||_1 + R_N (r+M) ||f||_\infty. \end{aligned}$$

puis, par récurrence, pour tout $f \in E^1$ et tout entier $k \geq 1$,

$$|||P_u^{kN}|||_1 \leq r^k |||f|||_1 + R_N \cdot \left(r^{k-1} + M \frac{1-r^{k-1}}{1-r}\right) ||f||_\infty.$$

En conséquence, la famille $\{P_u^{kN} f,\ k \geq 1\}$ est équicontinue et bornée dans $(E, ||\ ||_\infty)$. Cette dernière propriété, appliquée avec $f = P_u^m 1$, pour $m \in \{0, \ldots, N-1\}$, montre que $\{P_u^n 1,\ n \geq 1\}$, et donc $\{\frac{1}{n}\sum_{k=0}^{n-1} P_u^k 1,\ n \geq 1\}$, sont équicontinues et uniformément bornées dans $(E, ||\ ||_\infty)$. Grâce au théorème d'Ascoli, on peut extraire

de cette dernière suite une sous-suite convergeant vers un élément $\gamma \geq 0$ de E^1, 1-périodique, vérifiant par passage à la limite $\gamma(0) = 1$ et $P_u\gamma = \gamma$, ce qui prouve bien la propriété annoncée.

L'ensemble $Z(\gamma)$ des zéros de γ sur $[0,1]$ est un compact invariant pour m_0 (cf. §2). On a $\gamma(0) = \gamma(1) = 1$, et d'après (11), tout compact invariant contient 0 ou 1. Donc $\gamma > 0$.

Enfin, l'unicité de γ se démontre suivant les mêmes techniques que pour le théorème 2.5, en utilisant cette fois l'opérateur relativisé T défini sur E par $Tf = \gamma^{-1}P_u(\gamma f)$. Le lemme est ainsi démontré.

Preuve de (ii). Si $\hat{\phi} \in L^1(\mathbf{R})$, on a, en vertu du théorème 6.1, $\rho_u = 1$ et $\nu(\rho_u) = 1$. Par ailleurs, d'après le théorème 3.1, il existe une fonction $g_0 \in E^1$, positive ou nulle, telle que $P_a g_0 = \rho_a g_0 = \beta_1 g_0$. La fonction $g(x) = |\sin \pi x|^r \; g_0(x)$ appartient à E^1, et vérifie, d'après la formule $\sin 2x = 2 \sin x \cos x$,

$$P_u g(x) = 2^{-r} |\sin \pi x|^r \; P_a g_0(x) = 2^{-r} \beta_1 g(x),$$

d'où $2^{-r}\beta_1 \leq \rho_u$, soit $\beta_1 \leq 2^r$. Si $\beta_1 = 2^r$, alors $P_u g = g$, d'où, en vertu du lemme précédent, $g = g(0)\gamma$, soit $g \equiv 0$, ce qui est absurde.

Réciproquement, si $\beta_1 < 2^r$, le lemme 4.3 appliqué avec $w = a$ et $b = r$ prouve que $\hat{\phi} \in L^1(\mathbf{R})$ (on pourra s'inspirer de la méthode utilisée pour le théorème 4.2).

Calcul de s_1. Considérons un réel b tel que $b > \log_2 \beta_1$. On a

$$\begin{aligned}
\int_{-\infty}^{+\infty} |\lambda|^{r-b} |\hat{\phi}(\lambda)| d\lambda &= \int_{-\infty}^{+\infty} |\lambda|^{r-b} \prod_{k \geq 1} \left| \cos \frac{\pi\lambda}{2^k} \right|^r \prod_{k \geq 1} \left| v\left(\frac{\lambda}{2^k}\right) \right| d\lambda \\
&= \int_{-\infty}^{+\infty} |\lambda|^{r-b} \frac{|\sin \pi\lambda|^r}{|\pi\lambda|^r} \prod_{k \geq 1} \left| v\left(\frac{\lambda}{2^k}\right) \right| d\lambda \\
&\leq C \int_{-\infty}^{+\infty} \frac{1}{1 + |\lambda|^b} \prod_{k \geq 1} \left| v\left(\frac{\lambda}{2^k}\right) \right| d\lambda.
\end{aligned}$$

D'après le lemme 4.3, cette dernière intégrale est finie, d'où $s_1 \geq r - \log_2 \beta_1$. Procédons maintenant par l'absurde en supposant que $s_1 > r - \log_2 \beta_1$. Il existe alors un réel $b > r - \log_2 \beta_1$ tel que $\int_{-\infty}^{+\infty} |\lambda|^b |\hat{\phi}(\lambda)| d\lambda < \infty$. La fonction f, 1-périodique, positive ou nulle, définie par

$$f(x) = \sum_{k \in \mathbf{Z}} |x + k|^b |\hat{\phi}(x + k)|$$

est intégrable sur $[0,1]$. Remarquons que f est a priori à valeurs dans $[0, +\infty]$, et qu'elle est semi-continue inférieurement, comme limite croissante de fonctions continues.

Grâce à (3), il vient que

$$f(x) = 2^{-b} \sum_{k \in \mathbf{Z}} \left| \frac{x}{2} + \frac{k}{2} \right|^b u\left(\frac{x}{2} + \frac{k}{2}\right) \left| \hat{\phi}\left(\frac{x}{2} + \frac{k}{2}\right) \right|, \quad \forall x \in \mathbf{R}.$$

Découpant cette dernière somme suivant les indices k pairs et impairs, on obtient $P_u f = 2^{-b} f$, l'opérateur P_u étant considéré sur $L^1([0,1])$ dans l'égalité précédente.

LEMME 8.2. *L'ensemble des zéros de f est réduit à $\{0,1\}$.*

Preuve du lemme. Pour $x \in [0,1]$, nous appelons *trajectoire* de x pour m_0 tout sous-ensemble de $[0,1]$ de la forme $\{\sigma_n \cdots \sigma_1 x, n \geq 1\}$ où $\{\sigma_n, n \geq 1\}$ est une suite d'éléments de $\{S_0, S_1\}$ vérifiant $m_0(\sigma_n \cdots \sigma_1 x) \cdots m_0(\sigma_1 x) \neq 0$, pour tout $n \geq 1$. Les applications S_0 et S_1 ont été définies dans le §2. L'adhérence de l'ensemble des trajectoires de x est un compact invariant pour m_0 appelé *orbite* de x.

Commençons par prouver que, si f est nulle en un point x de $[0,1]$, alors f est identiquement nulle sur l'orbite O_x de x. Il est clair que f est identiquement nulle sur l'ensemble T_x des trajectoires de x. Soit $y \in O_x$: il existe une suite $(y_n)_{n\geq 1}$ d'éléments de T_x convergeant vers y. Comme f est semi-continue inférieurement, nous avons $0 \leq f(y) \leq \liminf_n f(y_n)$, d'où $f(y) = 0$, ce qui prouve la propriété annoncée.

Rappelons que $\hat{\phi}(k) = 0$, pour tout $k \in \mathbf{Z}$, $k \neq 0$. On a donc $f(0) = f(1) = 0$. Soit $x \neq 0, 1$: il reste à démontrer que $f(x) \neq 0$. A cet effet, on procède par l'absurde : si $f(x) = 0$, alors f est identiquement nulle sur l'orbite de x, qui par hypothèse contient 0 (ou 1). Il existe donc une suite $(x_n)_{n\geq 1}$ convergeant vers 0 (ou 1, mais la preuve est alors analogue) telle que $f(x_n) = 0$ pour tout $n \geq 1$. Ecrivant la relation $P_u f(x_n) = 2^{-b} f(x_n)$, et utilisant le fait que $\frac{1}{2}$ n'est pas un point d'accumulation de zéros de u, il vient que $f(\frac{x_n}{2} + \frac{1}{2}) = 0$ pour tout n assez grand, d'où, par passage à la "lim inf", $f(\frac{1}{2}) = 0$. On en déduit que f est identiquement nulle sur l'orbite de $\frac{1}{2}$, qui, d'après un résultat prouvé dans [18], coïncide avec $[0,1]$. Ainsi, $\hat{\phi}$ est identiquement nulle sur $\mathbf{R}$, ce qui évidemment est absurde. Le lemme est démontré.

Considérons maintenant la fonction h, positive ou nulle, 1-périodique, définie par $h(x) = |\sin \pi x|^{-r} f(x)$, si $x \in]0,1[$, et $h(0) = h(1) = (2^{r-b} - 1)^{-1} a(\frac{1}{2}) f(\frac{1}{2})$. La fonction h est a priori à valeurs dans $[0, +\infty]$. On vérifie aisément que h est semi-continue inférieurement. D'après le lemme 8.2, on a $h(x) \neq 0$, $\forall x \in [0,1]$. Donc $\inf_{x\in[0,1]} h(x) > 0$.

Pour $x \in [0,1]$, on pose $P_a h(x) = a(\frac{x}{2}) h(\frac{x}{2}) + a(\frac{x}{2} + \frac{1}{2}) h(\frac{x}{2} + \frac{1}{2})$. De l'identité $P_u f = 2^{-b} f$, et de la formule $\sin 2x = 2 \sin x \cos x$, il résulte que $P_a h(x) = 2^{r-b} h(x)$ pour tout $x \in]0,1[$. Par ailleurs, nous avons choisi $h(0)$ et $h(1)$ telles que $P_a h(0) = 2^{r-b} h(0)$ et $P_a h(1) = 2^{r-b} h(1)$, d'où finalement $P_a h(x) = 2^{r-b} h(x)$ pour tout $x \in [0,1]$.
Nous pouvons maintenant conclure : il existe une constante $d > 0$ telle que $\gamma(x) \leq d h(x)$, pour tout $x \in [0,1]$, où γ est la fonction du lemme 8.1. Donc pour tout $n \geq 1$, $\gamma(x) \leq d 2^{n(r-b)} \beta_1^{-n} h(x)$. Or, rappelons qu'on a supposé $b - \log_2 \beta_1$, soit $2^{r-b} < \beta_1$. Il vient que γ est identiquement nulle, ce qui est absurde. Le théorème est démontré.
□

Remerciments. Durant la préparation de ce travail, largement inspiré de [6] [7], j'ai pu bénéficier du soutien et des encouragements de Jean-Pierre Conze. Qu'il soit assuré de ma reconnaissance. Je tiens également à remercier Yves Meyer dont les conseils m'ont permis d'améliorer considérablement de cet article, et enfin Patrick Mear qui a calcule les coefficuents s_1 et s_2 des filtres de Butterworth (§6).

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative matrices in the mathematical science*, Computer Science and Applied Mathematics, Werner Rheinboldt, ed., 1979.

[2] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc. 93 (1991), pp. 1–186.

[3] A. Cohen, *Ondelettes, analyses multirésolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 7 (1990), pp. 439–459.

[4] A. COHEN, I. DAUBECHIES, AND J. C. FEAUVEAU, *Biorthogonal Bases of Compactly Supported Wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.
[5] A. COHEN AND I. DAUBECHIÈS, *Non-separable bidimensional wavelet bases*, Rev. Math. Iberoamericana, 9 (1993), pp. 51–137.
[6] J.-P. CONZE, *Sur la régularité des solutions d'une équation fonctionnelle*, Laboratoire de Probabilités, Université de Rennes I, Juin, 1989.
[7] J.-P. CONZE AND A. RAUGI, *Fonctions harmoniques pour un opérateur de transition et applications*, Bull. Soc. Math. France, 118 (1990), pp. 273–310.
[8] I. DAUBECHIES, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math., Vol. 61, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.
[9] I. DAUBECHIES AND J. C. LAGARIAS, *Two-scale difference equations. Part II. Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 24 (1992), pp. 1031–1079.
[10] G. DESLAURIERS AND S. DUBUC, *Interpolation dyadique*, dans Fractals, dimensions non entières et applications, Masson, Paris, 1987.
[11] G. DESLAURIERS, J. DUBOIS, AND S. DUBUC, *Multidimensional Iterative Interpolation*, Rapport Technique 41, Département de Mathematiques et Informatique, Univ. de Sherbrooke, 1988.
[12] N. DUNFORD AND J. T. SCHWARTZ, *Linear operator, Part. I*, Pure and Applied Mathematics, Vol VII, Wiley-Interscience, New York, 19xx.
[13] T. EIROLA, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.
[14] G. GRIPENBERG, *Unconditional bases of wavelets for Sobolev spaces*, SIAM J. Math. Anal., 24 (1993), pp. 1030–1042.
[15] H. HENNION, *Sur un théorème spectral et son application aux noyaux lipchitziens*, Proceeding of the Amer. Math. Soc., 118 (1993), pp. 627–634.
[16] C. HERLEY AND M. VETTERLI, *Wavelets and recursive filter banks*, IEEE Trans. Signal Processing, 41 (19xx), pp. 2536–2556.
[17] L. HERVÉ, *Régularité et conditions de bases de Riesz pour les fonctions d'échelle*, C.R. Acad. Sci., Paris, Ser. I, 335 (1992), pp. 1029–1032.
[18] ———, *Etude d'opérateurs quasi-compacts positifs. Applications aux opérateurs de transfert*, Ann. Inst. H. Poincaré, Probab. Statist., 30 (1994), pp. 437–466.
[19] ———, *Multi-Resolution Analysis of multiplicity d. Applications to dyadic interpolations*, Appl. Comput. Harmonic Anal., 1 (1994), pp. 299–315.
[20] ———, *Méthodes d'opérateurs quasi-compacts en analyse multirésolution, application à la construction de bases d'ondelettes et à l'interpolation*, Thèse, Laboratoire de Probabilités, Université de Rennes I, 1992.
[21] C. T. IONESCU-TULCEA AND G. MARINESCU, *Théorie ergodique pour une classe d'opérations non complétement continues*, Ann. Math., 52 (1950), pp. 140–147.
[22] M. KEANE, *Strongly Mixing g-Measures*, Inventiones Math., 16 (1972), pp. 309–324.
[23] S. MALLAT, *Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–88.
[24] J. L. MERRIEN, *A family of C^1 interpolants built by dichotomy*, Numer. Algorithms, 2 (1992), pp. 187–200.
[25] Y. MEYER, *Ondelettes et opérateurs* I, Hermann, Paris, 1990.
[26] O. RIOUL, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal., 23 (1992), pp. 1544–1576.
[27] L. F. VILLEMOES, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.
[28] ———, *Wavelet analysis of refinement equations*, SIAM J. Math. Anal., 25 (1994), pp. 1433–1460.

SIAM J. MATH. ANAL.
Vol. 26, No. 5, pp. 1386–1402, September 1995

016

SOME RESULTS ON THE CONVERGENCE OF SAMPLING SERIES BASED ON CONVOLUTION INTEGRALS*

SÔNIA M. GOMES† AND ELSA CORTINA‡

Abstract. A generalization of sampling series is introduced by considering expansions in terms of scaled translates of a basic function with coefficients given by sampled values of the convolution of a function f with a kernel of Fejér's type. Such expressions have been used in finite element approximations, sampling theory and, more recently, in wavelet analysis. This article is concerned with the convergence of these series for functions f that exhibit some kind of local singular behavior in time or frequency domains. Pointwise convergence at discontinuity points and Gibbs phenomena are analysed. The convergence in the H^s-norm is also investigated. Special attention is focused on multiresolution analysis approximations and examples using the Daubechies scaling functions are presented.

Key words. sampling series, wavelets, Gibbs phonomenon

AMS subject classifications. 41A25, 41A58

1. Introduction.

In this paper we shall deal with expansion series of the form

$$\sum_{k=-\infty}^{\infty} c_{h,k}\Phi(h^{-1}x-k), \tag{1.1}$$

in terms of the scaled translates of a basic function Φ. When the coefficients are the sampled values of a function f, $c_{h,k}=f(hk)$, this representation is denoted by

$$S_hf(x)=\sum_{k=-\infty}^{\infty} f(hk)\Phi(h^{-1}x-k), \tag{1.2}$$

and it is known in sampling theory as sampling series. An historical overview of this matter is given in [4].

Our aim here is to study expansions of type (1.1) with coefficients given by the sampled values of $(f*\mu_h)(hk)$ of the convolution of f with a kernel of Fejér's type $\mu_h(v)=h^{-1}\mu(h^{-1}v)$. That is,

$$Q_hf(x)=S_h(R_hf)(x)=\sum_{-\infty}^{\infty} R_hf(hk)\Phi(h^{-1}x-k), \tag{1.3}$$

where

$$R_hf(x)=\int_{-\infty}^{\infty} f(x-v)h^{-1}\mu(h^{-1}v)dv, \tag{1.4}$$

* Received by the editors March 31, 1993; accepted for publication (in revised form) January 27, 1994. This work was partially supported by Consejo Nacional de Investigaciones Científicas y Técnicas–Argentina and Conselho Nacional de Desenvolvimento Científico e Tecnológico–Brasil.

† Institúto de Matemática, Universidade de Campinas, Caixa Postal 6065, 13081-970 Campinas, São Paulo, Brasil. The work of this author was partially supported by Fundação de Amparo à Pesquisa do Estado de São Paulo, Brasil.

‡ Dirección General de Investigación y Desarrollo, Avda. Corrientes 1516, 1042 Buenos Aires, Argentina, Research Fellow of the Consejo Nacional de Investigaciones Científicas y Técnicas. The work of this author was partially supported by European Commission contract/grant CTI* CT 91-0944.

with $\mu \in L^1(\mathbf{R})$ and $\int_{-\infty}^{\infty} \mu(v)dv = 1$. Observe that sampling series correspond to the limit case when μ is the delta distribution.

Expansions of type (1.3) have been extensively used in finite element approximations, where typically μ and Φ have compact support and S_h and R_h are called prolongation and restriction operators, respectively [1].

Multiresolution analysis of $L^2(\mathbf{R})$ (cf. [17], [18], and [9]) is another interesting context in which expansions (1.3) appear. In a multiresolution analysis of $L^2(\mathbf{R})$ a function $f \in L^2(\mathbf{R})$ can be decomposed as follows:

$$f(x) = \sum_{k\in\mathbf{Z}} \langle f, \Phi_{jk}\rangle \Phi_{jk}(x) + \sum_{\nu\geq j} \sum_{k\in\mathbf{Z}} \langle f, \Psi_{\nu k}\rangle \Psi_{\nu k}(x). \tag{1.5}$$

The function $\Phi(x)$ appearing in the first term in the right-hand side of the above expression is called *scaling function*. For $j \in \mathbf{Z}$ the sets

$$\{\Phi_{jk}(x) = 2^{j/2}\Phi(2^j x - k), k \in \mathbf{Z}\}$$

form orthonormal bases of embedded closed subspaces $V_j \subset L^2(\mathbf{R})$ and

$$\Pi_j f(x) = \sum_{k\in\mathbf{Z}} \langle f, \Phi_{jk}\rangle \Phi_{jk}(x) \tag{1.6}$$

is the orthogonal projection of f onto V_j. Observe that $\Pi_j f$ has the form (1.1), where the coefficients are the L^2-scalar products $\langle f, \Phi_{jk}\rangle 2^{j/2}$, and can also be represented in the convolution form (1.3) with $h = 2^{-j}$ and $\mu(v) = \Phi(-v)$. Similarly, the functions $\Psi_{\nu k}(x)$ in the second term of (1.5) are defined as

$$\Psi_{\nu k}(x) = 2^{\nu/2}\Psi(2^\nu x - k)$$

in terms of the function $\Psi(x)$, which is usually called *basic wavelet*, and the set $\{\Psi_{\nu k}(x), \nu, k \in \mathbf{R}\}$ constitutes an orthonormal basis for $L^2(\mathbf{R})$. Moreover, for every j, the closed subspace W_j spanned by $\{\Psi_{jk}(x), k \in \mathbf{Z}\}$ is the orthogonal complement of V_j in V_{j+1}. Consequently, in this kind of decomposition the higher resolution approximation $\Pi_{j+1} f$ is obtained by adding to $\Pi_j f$ a *high frequency* component

$$D_j f = \sum_{k\in\mathbf{Z}} \langle f, \Psi_{jk}\rangle \Psi_{jk}(x)$$

corresponding to the orthogonal projection of f onto W_j. A multiresolution expansion (1.5) is a discrete version of the *continuous wavelet transform*

$$Wf(a,b) = a^{-1/2} \int_{-\infty}^{\infty} f(x)\Psi\left(\frac{x-b}{a}\right) dx.$$

This technique provides an adequate framework to analyse those phenomena that are well localized in time or frequency domains. It has received considerable attention in the last years and had been succesfully applied in several fields of mathematics, physics, and signal analysis (cf. [6], [14], and [19]).

There also exist generalizations of expansions (1.5) in the form

$$f(x) = \sum_{k\in\mathbf{Z}} \langle f, \Phi^*_{jk}\rangle \Phi_{jk}(x) + \sum_{\nu\geq j} \sum_{k\in\mathbf{Z}} \langle f, \Psi^*_{\nu k}\rangle \Psi_{\nu k}(x)$$

that differ in that the set $\{\Phi_{jk}, \Psi_{jk}\}$ of the *synthesizing functions* is not orthogonal and the *analysing functions* $\{\Phi^*_{jk}, \Psi^*_{jk}\}$ are not necessarily the same as the synthesizing functions. The projectors

$$P_j f(x) = \sum_{k \in \mathbf{Z}} \langle f, \Phi^*_{jk} \rangle \Phi_{jk}(x) \tag{1.7}$$

also have the form (1.3) with $h = 2^{-j}$ and $\mu(v) = \Phi^*(-v)$. This kind of decomposition includes the *phi-transform* [13] and biorthogonal multiresolution expansions [5], [22], all of them being appropiate for local time-frequency analysis.

With this motivation, most of the results of the present paper are concerned with the convergence of series (1.3) for functions f that exhibit some kind of local singular behavior in the time or frequency domain. In the examples presented here we used the compactly supported scaling functions $\Phi = \Phi_N$ constructed by I. Daubechies in [8]. Even though the emphasis is in the context of orthogonal multiresolution analysis, we also indicate how the results can be stated for a wider class of functions Φ and μ. Section 2 is mostly devoted to the pointwise convergence at jump discontinuities and to Gibbs phenomena. For instance, we will deduce that in multiresolution approximations $\Pi_j f$ the Gibbs phenomenon does occur near a discontinuity point but, due to the local support of Φ, it is restricted to a small neighborhood. In §3 we analyse the convergence in the H^s-norm. A classical result from finite elements theory states that the order of convergence depends on the Strang–Fix condition and on a moment relation. Here this result is obtained for basic functions Φ and μ that do not necessarily have compact support.

2. Pointwise convergence. In this section we will investigate the pointwise convergence of expansions (1.3). We are mainly interested in the analysis of the Gibbs phenomenon at jump discontinuities. To carry out this analysis we need first to state the pointwise and uniform convergence of $Q_h f(x)$ at continuity points of f. When asking for conditions on Φ and μ such that

$$\lim_{h \to 0} Q_h f(x) = f(x)$$

at each point of continuity of f, we are led to consider the same question for the operators S_h and R_h.

Concerning the convergence of singular integrals (1.4) we refer to Proposition 3.2.1 in [2]. It states that, if $\mu \in L^1(\mathbf{R})$ and $\int_{-\infty}^{\infty} \mu(v)dv = 1$, and $f \in L^\infty(\mathbf{R})$ is continuous at t, then $R_h f(t) \to f(t)$, as $h \to 0$. Furthermore, if f is continuous on $(a - \eta, b + \eta)$ for some $\eta > 0$, $a < b, a, b \in \mathbf{R}$, this convergence is uniform for all $t \in [a, b]$.

Let us also recall some results for sampling series (1.2) given in [3]. For a bounded function Φ such that $\sum_{k \in \mathbf{Z}} |\Phi(x - k)|$ converges uniformly on $[0, 1)$, it is proved in [3, Thm. 1], that the following assertions are equivalent:

(a) for each $f \in L^\infty(\mathbf{R})$ and each point t of continuity of f, $S_h f(t) \to f(t)$, as $h \to 0$.

(b) Φ satisfies the partition of the unity property

$$\sum_{k \in \mathbf{Z}} \Phi(x - k) = 1, \tag{2.1}$$

for all $x \in [0, 1]$.

In what follows NL^1 denotes the set of those functions $\mu \in L^1(\mathbf{R})$ normalized by $\int_{-\infty}^{\infty} \mu(v)dv = 1$. We call Nl^1 the set of the bounded functions $\Phi(x)$ such that $\sum_{k\in\mathbf{Z}} |\Phi(x-k)|$ converges uniformly on $[0,1)$ and that satisfy (2.1).

Combining the above results we can easily prove the following theorem.

THEOREM 2.1. *Assume $\mu \in NL^1$ and $\Phi \in Nl^1$. Then, for each $f \in L^\infty(\mathbf{R})$ and each point $t \in \mathbf{R}$ of continuity of f,*

$$\lim_{h\to 0} Q_h f(t) = f(t).$$

Moreover, if f is continuous on $(a-\eta, b+\eta)$ for some $\eta > 0$, $a < b, a, b \in \mathbf{R}$, the convergence is uniform on $[a,b]$.

Note that, if $\Phi \in L^1(\mathbf{R})$, the partition of the unity property implies that

$$\widehat{\Phi}(2k\pi) = \begin{cases} 1, & k = 0, \\ 0, & k \in \mathbf{Z} \setminus \{0\}, \end{cases}$$

where

$$\widehat{\Phi}(\xi) = \int_{-\infty}^{\infty} \Phi(x) e^{-i\xi x} dx$$

is the Fourier transform of $\Phi(x)$. There exist additional conditions that guarantee that the converse is also true (e.g., if Φ is continuous).

A function Φ is called r-regular if it is r times differentiable and for all indices β, such that $0 \le \beta \le r$,

$$|d^\beta \Phi / dx^\beta| \le C_n (1+|x|)^{-n}$$

for all integers $n \ge 0$.

According to Y. Meyer [18], r-regular scaling functions corresponding to multiresolution analysis of $L^2(\mathbf{R})$ are in Nl^1, and all the zeros of $\widehat{\Phi}$ at $\xi = 2k\pi, k \ne 0$, have at least order r+1. Therefore, as a consequence of Theorem 2.1 we have the following corollary.

COROLLARY 2.2. *Let Φ be an r-regular scaling function and $\Pi_j f$ the operator defined in* (1.6). *If $f \in L^\infty(\mathbf{R})$ is continuous at t, then $\Pi_j f(t) \to f(t)$. Moreover, if f is continuous on $(a-\eta, b+\eta)$ for some $\eta > 0$, $a < b, a, b \in \mathbf{R}$, then*

$$\lim_{j\to\infty} \Pi_j f(x) = f(x)$$

uniformly on $[a,b]$

2.1. Convergence at discontinuity points. Next we will study the behavior of the series (1.3) for functions f that have a jump discontinuity at a point t, where the limits

$$f(t^+) = \lim_{\varepsilon\to 0+} f(t+\varepsilon),$$
$$f(t^-) = \lim_{\varepsilon\to 0+} f(t-\varepsilon)$$

exist and are different. Consider first a simple example.

Example 2.3. Let $f(x) = H(x-t)$, where H is the Heavyside function $H(x) = 0$ for $x < 0$ and $H(x) = 1$ for $x \ge 0$.

$$R_h f(hk) = \int_{-\infty}^{k-h^{-1}t} \mu(v)dv$$

and

$$Q_h f(t+x) = \sum_{k\in\mathbf{Z}} \Phi(h^{-1}(t+x)-k) \int_{-\infty}^{k-h^{-1}t} \mu(v)dv. \tag{2.2}$$

Looking at the above expression, we are naturally led to define

$$\gamma(x,y) = \Phi(x) \int_{-\infty}^{-y} \mu(v)dv$$

and

$$G(u,w) = \sum_{k\in\mathbf{Z}} \gamma(u+w-k, w-k). \tag{2.3}$$

Therefore, (2.2) can be written as

$$Q_h f(t+x) = G(h^{-1}x, h^{-1}t),$$

and at the discontinuity point one has

$$Q_h f(t) = G(0, h^{-1}t).$$

For $\mu \in NL^1$ and $\Phi \in Nl^1$, the series (2.3) converges for all u and $w \in \mathbf{R}$, and represents a function that satisfies $G(u, w+1) = G(u,w)$. The function $\Gamma(w)$ defined by

$$\Gamma(w) = G(0,w)$$

is therefore a 1-periodic function. Using this notation in the above example, we can write $Q_h f(t) = \Gamma(h^{-1}t)$.

Notice that, in the limit case $\mu(v) = \delta(v-0)$, $\Gamma(w) = \sum_{w<k}^{\infty} \Phi(w-k)$. It coincides with the function $\Psi_{\Phi}^{-}(w)$ introduced in [3], where the theorem that follows is proved for the particular case of sampling series.

THEOREM 2.4. *Let $f \in L^{\infty}(\mathbf{R})$ have a jump at t, and $\alpha \in \mathbf{R}$. Then for $\mu \in NL^1$ and $\Phi \in Nl^1$, the following two assertions are equivalent:*

(i) $\lim_{h\to 0} Q_h f(t) = \alpha f(t^+) + (1-\alpha) f(t^-)$.

(ii) $\Gamma(x) = \alpha$, $x \in [0,1)$.

Proof. Let us define

$$g_t(x) = \begin{cases} f(x) - f(t^-), & x < t, \\ 0, & x = t, \\ f(x) - f(t^+), & x > t. \end{cases} \tag{2.4}$$

Then $g_t \in L^{\infty}(\mathbf{R})$ is continuous at t and

$$R_h g_t(x) = \int_{-\infty}^{\infty} g_t(x-hv)\mu(v)dv = R_h f(x) - f(t^-) \int_{h^{-1}(x-t)}^{\infty} \mu(v)dv$$
$$- f(t^+) \int_{-\infty}^{h^{-1}(x-t)} \mu(v)dv.$$

Calling

$$I_{t,h}(x) = \int_{-\infty}^{h^{-1}(x-t)} \mu(v)dv,$$

we have

$$R_h g_t(hk) = R_h f(hk) - I_{t,h}(hk) f(t^+) - [1 - I_{t,h}(hk)] f(t^-).$$

Consequently

$$\begin{aligned} Q_h f(x) = Q_h g_t(x) + f(t^+) \sum_{k=-\infty}^{\infty} I_{t,h}(hk)\Phi(h^{-1}x - k) \\ + f(t^-) \left[1 - \sum_{k=-\infty}^{\infty} I_{t,h}(hk)\Phi(h^{-1}x - k) \right]. \end{aligned} \tag{2.5}$$

Note that

$$\sum_{k=-\infty}^{\infty} I_{t,h}(hk)\Phi(h^{-1}t - k) = \sum_{k=-\infty}^{\infty} \Phi(h^{-1}t - k) \int_{-\infty}^{k-h^{-1}t} \mu(v)dv = \Gamma(h^{-1}t).$$

Therefore

$$Q_h f(t) = Q_h g_t(t) + \Gamma(h^{-1}t) f(t^+) + [1 - \Gamma(h^{-1}t)] f(t^-). \tag{2.6}$$

Now we use Theorem 2.1, which ensures that $\lim_{h\to 0} Q_h g_t(t) = 0$. Hence (ii) implies

$$\lim_{h\to 0} Q_h f(t) = \alpha f(t^+) + (1 - \alpha) f(t^-).$$

Conversely, if (i) holds, from equation (2.6) it follows that

$$\lim_{h\to 0} \Gamma(h^{-1}t) = \alpha$$

and assertion (ii) is obvious. □

Because $\Gamma(w)$ is hardly ever a constant, $\lim_{h\to 0} Q_h f(t)$ does not exist in general. From equation (2.6) we see that $Q_h f(t)$ will oscillate between the values

$$\lambda_1 = \alpha_1 f(t^+) + [1 - \alpha_1] f(t^-)$$

and

$$\lambda_2 = \alpha_2 f(t^+) + [1 - \alpha_2] f(t^-),$$

where $\alpha_1 = \limsup \Gamma(w)$ and $\alpha_2 = \liminf \Gamma(w), w \in [0, 1)$. Nevertheless, specific discrete scale parameters h can be chosen such that $\lim_{h\to 0} Q_h f(t)$ does exist. For example, choose h such that the differences $h^{-1}t - [h^{-1}t]$ converge to some $w_0 \in [0, 1)$ as $h \to 0$. If Γ is continuous, then

$$\lim_{h\to 0} Q_h f(t) = f(t^+)\Gamma(w_0) + f(t^-)[1 - \Gamma(w_0)].$$

This is precisely what happens in multiresolution approximations $\Pi_j f$ of functions that have jump discontinuities at dyadic points $t = m2^s, m, s \in \mathbf{Z}$.

COROLLARY 2.5. *Let $f \in L^\infty(\mathbf{R})$ have a jump at t. If t is a dyadic point, then*

$$\Pi_j f(t) \to f(t^+)\Gamma_\Phi(0) + f(t^-)[1 - \Gamma_\Phi(0)]$$

as $h = 2^{-j} \to 0$, where

$$\Gamma_\Phi(w) = \sum_{k \in \mathbf{Z}} \Phi(w-k) \int_{w-k}^{\infty} \Phi(v)dv.$$

Figure 2.1 illustrates the functions Γ_Φ for Daubechies's scaling functions Φ_2 and Φ_4 which are not constant.

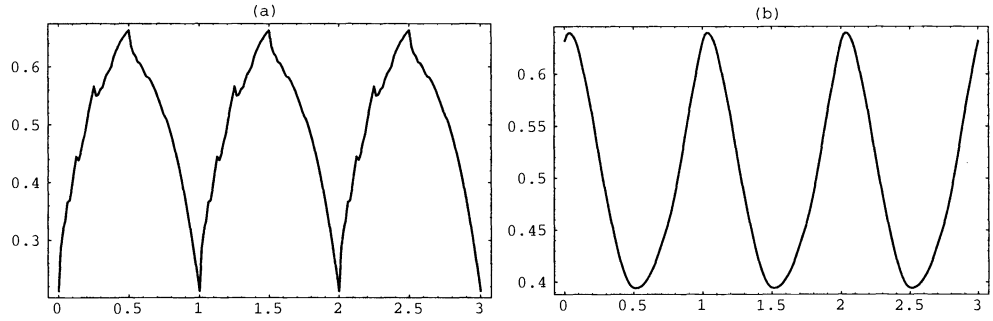


FIG. 2.1. *The function $\Gamma_\Phi(x)$ for $\Phi = \Phi_N$;* (a) *$N = 2$ and* (b) *$N = 4$.*

2.2. The Gibbs phenomenon. The convergence of the most popular expansion series of orthogonal functions, e.g., Fourier, Legendre, or Chebyshev series, is nonuniform in the neighborhood of discontinuity points of f where strong oscillations appear. This nonuniform behavior is called the *Gibbs phenomenon*, and it generally affects the rate of convergence even far away from the point of discontinuity where f is smooth. We will analyse here the Gibbs phenomenon in approximations $Q_h f$.

Considering again the translated Heavyside function $f(x) = H(x-t)$, we can see that, for all $x \in \mathbf{R}$,

$$Q_h f(t + hx) = G(x, h^{-1}t), \tag{2.7}$$

The Gibbs phenomenon will occur if G exhibits over ($G > 1$) or undershoots ($G < 0$). For example, let $t = 0$ and suppose $G(x,0) > 1$ for some $x \geq 0$. Then, $t_h = hx \to 0$ as $h \to 0$ and $Q_h f(t_h) = G(x,0) > f(0^+)$.

The result in (2.7) can be generalized as follows.

THEOREM 2.6. *Let $\Phi \in Nl^1$, $\mu \in NL^1$, and $f \in L^\infty(\mathbf{R})$, where f has an isolated jump discontinuity at t. Then, as $h \to 0$,*

$$Q_h f(t + hx) - f(t^+)G(x, h^{-1}t) + f(t^-)[1 - G(x, h^{-1}t)] \to 0 \tag{2.8}$$

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

Proof. Define g_t as in (2.4). From (2.5) it follows that

$$Q_h f(t+hx) = Q_h g_t(t+hx) + f(t^+) \sum_{k\in\mathbf{Z}} I_{t,h}(hk)\Phi(x+h^{-1}t-k)$$
$$+ f(t^-)\left[1 - \sum_{k\in\mathbf{Z}} I_{t,h}(hk)\Phi(x+h^{-1}t-k)\right],$$

where

$$\sum_{k\in\mathbf{Z}} I_{t,h}(hk)\Phi(x+h^{-1}t-k) = \sum_{k\in\mathbf{Z}} \Phi(x+h^{-1}t-k)\int_{-\infty}^{k-h^{-1}t} \mu(v)dv = G(x,h^{-1}t).$$

Therefore

$$Q_h f(t+hx) = Q_h g_t(t+hx) + f(t^+)G(x,h^{-1}t) + f(t^-)[1-G(x,h^{-1}t)].$$

Observe that $g_t(t) = 0$ and $g_t(x)$ is continuous in a neighborhood of $x = t$. The result on uniform convergence of Theorem 2.1 implies that $Q_h g_t(t+hx) \to 0$ as $h \to 0$. The proof is hereby complete. □

We conclude from this theorem that the convergence behavior of $Q_h f(t+hx)$ near the discontinuity point t is determined by the values $G(x,w_h)$, where $w_h = h^{-1}t - [h^{-1}t]$. Equation (2.8) ensures that $Q_h f(t+hx)$ will range between the values

$$m(x)f(t^+) + [1-m(x)]f(t^-)$$

and

$$M(x)f(t^+) + [1-M(x)]f(t^-),$$

where $m(x) = \liminf_{h\to 0} G(x,w_h)$ and $M(x) = \limsup_{h\to 0} G(x,w_h)$. If $M(x) > 1$ or $m(x) < 0$ for some $x \in \mathbf{R}$, then the Gibbs phenomenon appears. Since $G(u,v)$ characterizes the occurrence of the Gibbs phenomenon, we call it the *Gibbs function.*

Remark 2.7. A result similar to the above theorem was obtained in [20] for approximations by periodic spline functions. Here the Gibbs function G plays the role of the sine integral function $\mathrm{Si}(x)$ for trigonometrical series and of Gibbs spline $S_{[k]}(x)$ for spline approximations.

Example 2.8. In this example we apply the result of the above theorem to multiresolution analysis approximations $\Pi_j f$. The corresponding Gibbs function is

$$G_\Phi(u,w) = \sum_{k\in\mathbf{Z}} \Phi(u+w-k)\int_{w-k}^{\infty} \Phi(v)dv.$$

If $f \in L^\infty(\mathbf{R})$ has a jump at a dyadic point t, from Theorem 2.6 we conclude that

$$\Pi_j f(t+2^{-j}x) \to f(t^+)G_\Phi(x,0) + f(t^-)[1-G_\Phi(x,0)]$$

as $j \to \infty$.

Figures 2.2(a)–2.4(a) show the graphs of $G_{\Phi_N}(x,0)$ for Daubechies's scaling functions $\Phi = \Phi_N$, $N = 2,3,4$, where over- and undershoots can be seen. Since the Φ_N are supported on $[0,2N-1]$, it is easy to prove that $G_{\Phi_N}(x,0) = 0$ for $x \le 2-2N$

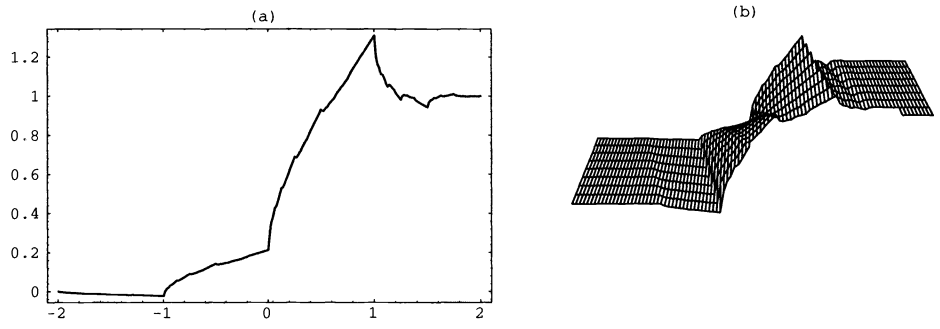


FIG. 2.2. *The Gibbs function* $G_\Phi(x,w)$ *for* $\Phi = \Phi_2$; (a) $w = 0$ *and* (b) $w \in [0,1]$.

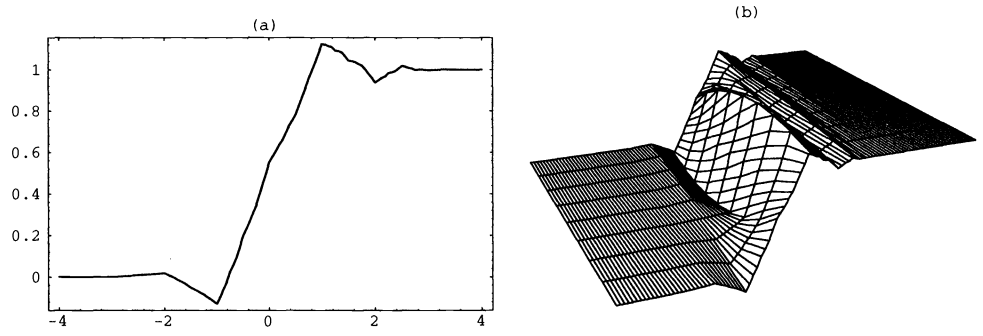


FIG. 2.3. *The Gibbs function* $G_\Phi(x,w)$ *for* $\Phi = \Phi_3$; (a) $w = 0$ *and* (b) $w \in [0,1]$.

and $G_{\Phi_N}(x,0) = 1$ for $x \geq 2N-2$. This means that, for these cases, the Gibbs phenomenon does occur in approximations $\Pi_j f$ of a function f having a jump discontinuity at a dyadic rational. However, it is localized in a neighborhood of the discontinuity point. In Table 2.1 the values of $G_{\Phi_N}(0,0) = \Gamma_{\Phi_N}(0)$, $N = 2,\ldots,6$, the minimum $G_{\Phi_N}(\alpha,0)$, and the maximum $G_{\Phi_N}(\beta,0)$ are displayed.

If the discontinuity point t is not dyadic, $\Pi_j f(t + 2^{-j}x)$ will diverge in general, but the Gibbs phenomenon will also occur. Figures 2.2(b)–2.4(b) show the graphs of $G_{\Phi_N}(x,w)$, $w \in [0,1]$. They show over- and undershoots whose positions and amplitudes change with w. Observe that, for each $w \in (0,1)$, $G_{\Phi_N}(x,w) = 0$ for $x \leq 1-2N$ and $G_{\Phi_N}(x,w) = 1$ for $x \geq 2N-1$.

Remark 2.9. *The Gibbs phenomenon of higher order in biorthogonal multiresolution approximations.* The analysis of a Gibbs phenomenon of higher order, say for a function f with a jump discontinuity in its derivative $\frac{d^k}{dx^k} f$, for some $k \geq 1$, may be

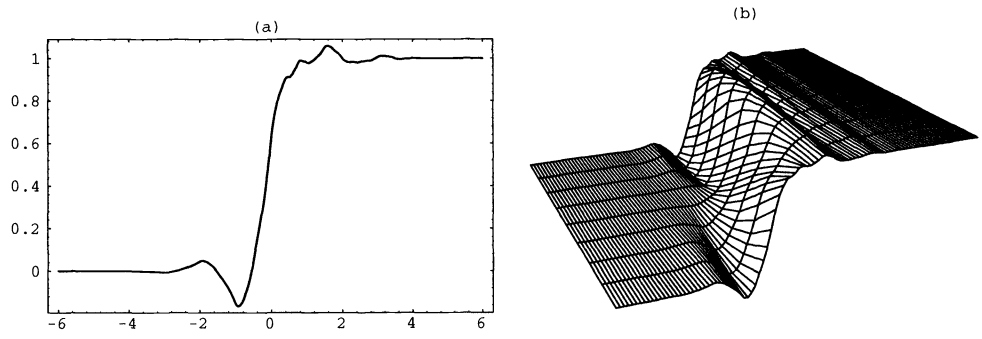


FIG. 2.4. *The Gibbs function* $G_\Phi(x,w)$ *for* $\Phi = \Phi_4$; (a) $w = 0$ *and* (b) $w \in [0,1]$.

TABLE 2.1

N	$G_{\Phi_N}(0,0)$	α	$G_{\Phi_N}(\alpha,0)$	β	$G_{\Phi_N}(\beta,0)$
2	0.2113248	-1.0000000	-0.0223290	1.0000000	1.3110042
3	0.5522790	-1.0000000	-0.1299706	1.0234375	1.1241512
4	0.6315820	-0.9453125	-0.1660389	1.5937500	1.0591883
5	0.4816549	-0.7890625	-0.0936842	0.8281250	1.1136424
6	0.4050844	-1.4531250	-0.0157200	0.9609375	1.1635085

reduced to study the problem at $k = 0$. For example, consider a biorthogonal multiresolution analysis $\{V_j, V_j^*\}_{j\in\mathbf{Z}}$ with dual compactly supported scaling functions Φ and Φ^* in $C^\epsilon, \epsilon > 0$. As pointed out in [16], if $\Phi \in C^{1+\epsilon}$, there exists another biorthogonal multiresolution analysis $\{\widetilde{V}_j, \widetilde{V}_j^*\}$ such that

$$\frac{d}{dx} \circ P_j = \widetilde{P}_j \circ \frac{d}{dx},$$

where P_j and $\widetilde{P}_j$ are the corresponding projectors operators (1.7). This commutation formula shows that a Gibbs phenomenon of first order will occur for P_j if and only if a Gibbs phenomenon of order zero appears in approximations by $\widetilde{P}_j f$. This procedure is straightforward for higher orders ($k > 1$).

3. Convergence in $H^s(\mathbf{R})$. In this section we will study how accurately functions $f \in H^s(\mathbf{R})$ can be approximated by expansions (1.1) in the H^s-norm. In the context of finite element approximations the answer to this question is a classical result obtained by Strang and Fix [21]. Assuming that $\Phi \in H^m(\mathbf{R})$ and has compact support, they established that *smooth functions can be approximated by expansions* (1.3) *with error* $O(h^{m+1-s})$ *in the* H^s*-norm,* $s \leq m$, *if and only if the polynomials of degree* $\leq m$ *can be written as linear combinations of* Φ *and its integer translates.* If Φ is normalized by $\int \Phi(x)dx = 1$, this is equivalent to the following hypothesis.

HYPOTHESIS 3.1. $\Phi \in Nl^1$, *and the zeros of its Fourier transform* $\widehat{\Phi}$ *at all the points* $\xi = 2\pi j, j \neq 0$, *are of order* $m + 1$.

This condition is sometimes referred to as the Strang–Fix condition.

In general, additional assumptions are necessary to get good accuracy in arbitrary approximations (1.1). For instance, for expansions (1.3)–(1.4), the following moment relation is required.

HYPOTHESIS 3.2. *μ and Φ are related by*

$$\widehat{\Phi}(\xi)\widehat{\mu}(\xi) = 1 + O(\xi^{q+1}), \tag{3.1}$$

$q \geq 0$.

For compactly supported $\mu \in NL^1 \cap L^\infty(\mathbf{R})$ and $\Phi \in H^r(\mathbf{R})$, and for $f \in H^{p+1}(\mathbf{R})$, with $p = \min\{m, q\}$, the above hypotheses imply that

$$\|f - Q_h f\|_{H^s} \leq Ch^{p+1-s}\|f\|_{H^{p+1}}, \tag{3.2}$$

where $0 \leq s \leq \min\{p, r\}$ and the constant $C = C(s)$ is independent of f (see, for example, [1] and also [7]).

The following example illustrates how these requirements are fulfilled by the scaling functions of Daubechies.

Example 3.3. The scaling functions of Daubechies $\Phi = \Phi_N$ satisfy the Strang–Fix condition with $m = N - 1$ (cf. [8]). Note that their Sobolev indices $r = r(N)$ are much less than $N - 1$ (cf. [10]). Let us look at the moment relation for the orthogonal projection $\Pi_j f$ which corresponds to $\mu(x) = \Phi(-x)$. Orthonormality implies that

$$\sum_{k \in \mathbf{Z}} |\widehat{\Phi}(\xi + 2k\pi)|^2 = 1$$

for all $\xi \in \mathbf{R}$. Consequently

$$\widehat{\Phi}(\xi)\widehat{\mu}(\xi) = |\widehat{\Phi}(\xi)|^2 = 1 - \sum_{k \neq 0} |\widehat{\Phi}(\xi + 2k\pi)|^2 = 1 + O(|\xi|^{2N}). \tag{3.3}$$

Therefore, if $f \in H^N(\mathbf{R})$ and $0 \leq s \leq r(N)$,

$$\|f - \Pi_j f\|_{H^s} \leq C2^{-j(N-s)}\|f\|_{H^N}. \tag{3.4}$$

Next we shall see that it is possible to have the convergence estimate (3.2) under weaker conditions, i.e., without assuming μ and Φ with compact support.

THEOREM 3.4. *Assume that $\mu \in NL^1$, Φ is r-regular and satisfies the Strang–Fix condition, and they are related by* (3.1). *Let $p = \min\{m, q\}$. Then the error estimate* (3.2) *holds for all $f \in H^{p+1}(\mathbf{R})$.*

Proof. The proof can be carried out in a way analogous to [21, Thm. I]. In order to estimate

$$\|f - Q_h f\|_{H^s}^2 = \int_{-\infty}^{\infty} (1 + |\xi|^2)^s |\widehat{f}(\xi) - \widehat{Q_h f}(\xi)|^2 d\xi$$

we have to show that the integrals

$$I_1 = \int_{|\xi| \leq \pi/h} (1 + |\xi|^2)^s |\widehat{f}(\xi) - \widehat{Q_h f}(\xi)|^2 d\xi$$

and

$$I_2 = \int_{|\xi| \geq \pi/h} (1 + |\xi|^2)^s |\widehat{f}(\xi) - \widehat{Q_h f}(\xi)|^2 d\xi$$

are all bounded by $Ch^{2(p+1-s)}\|f\|^2_{H^{p+1}}$.

The Fourier transform of $Q_hf(x)$ is

$$\widehat{Q_hf}(\xi) = h\widetilde{R_hf}(h\xi)\widehat{\Phi}(h\xi), \tag{3.5}$$

where

$$\widetilde{R_hf}(\xi) = \sum_k R_hf(hk)e^{-ik\xi} = h^{-1}\sum_k \widehat{R_hf}(h^{-1}(\xi + 2k\pi)).$$

Therefore,

$$\widehat{Q_hf}(\xi) = \widehat{f}(\xi)\widehat{\mu}(h\xi)\widehat{\Phi}(h\xi) + \sum_{k\neq 0}\widehat{R_hf}(h^{-1}(h\xi + 2k\pi))\widehat{\Phi}(h\xi).$$

Replacing this expression in I_1 one has

$$\begin{aligned} I_1 &\leq 2\Bigg\{y\int_{|\xi|\leq\pi/h}(1+|\xi|^2)^s|\widehat{f}(\xi)[1-\widehat{\mu}(h\xi)\widehat{\Phi}(h\xi)]|^2d\xi \\ &\qquad + \int_{|\xi|\leq\pi/h}(1+|\xi|^2)^s|\sum_{j\neq 0}\widehat{R_hf}(\xi+2\pi jh^{-1})|^2|\widehat{\Phi}(h\xi)|^2d\xi\Bigg\} \\ &= 2\{(\mathrm{a})+(\mathrm{b})\}. \end{aligned}$$

The moment relation (3.1) implies that

$$(\mathrm{a}) \leq Ch^{2(q+1-s)}\int_{-\infty}^{\infty}(1+|\xi|^2)^{q+1}|\widehat{f}(\xi)|^2d\xi$$

and, by the Cauchy–Schwartz inequality,

$$(\mathrm{b}) \leq Ch^{2(q+1-s)}\int_{-\infty}^{\infty}|\xi|^{2(q+1)}|\widehat{f}(\xi)|^2d\xi.$$

I_2 is also split into two terms:

$$\begin{aligned} I_2 &\leq 2\left\{\int_{|\xi|\geq\pi/h}(1+|\xi|^2)^s|\widehat{f}(\xi)|^2d\xi + \int_{|\xi|\geq\pi/h}(1+|\xi|^2)^s|\widehat{Q_hf}(\xi)|^2d\xi\right\} \\ &= 2\{(\mathrm{i})+(\mathrm{ii})\}. \end{aligned}$$

Notice that

$$\begin{aligned} (\mathrm{i}) &\leq C\int_{|\xi|\geq\pi/h}|h\xi|^{2(p+1-s)}(1+|\xi|^2)^s|\widehat{f}(\xi)|^2d\xi \\ &\leq Ch^{2(p+1-s)}\int_{-\infty}^{\infty}(1+|\xi|^2)^{2(p+1)}|\widehat{f}(\xi)|^2d\xi. \end{aligned}$$

Replacing (3.5) in (ii) we get

$$\begin{aligned} (\mathrm{ii}) &= h^2\int_{|\xi|\geq\pi/h}(1+|\xi|^2)^s|\widetilde{R_hf}(h\xi)|^2|\widehat{\Phi}(h\xi)|^2d\xi \\ &\leq Ch^{1-2s}\int_{-\pi}^{\pi}|\widetilde{R_hf}(\xi)|^2\sum_{j\neq 0}|\widehat{\Phi}(\xi+2j\pi)|^2|\xi+2j\pi|^{2s}d\xi. \end{aligned}$$

Suppose that, for $0 \le s \le r$,

$$\sum_{j\neq 0} |\widehat{\Phi}(\xi + 2j\pi)|^2 |\xi + 2j\pi|^{2s} = O(|\xi|)^{2m+2} \tag{3.6}$$

as $\xi \to 0$. From this estimate we hence conclude

$$\begin{aligned}\text{(ii)} &\le Ch^{1-2s} \int_{-\pi}^{\pi} |\widetilde{R_h f}(\xi)|^2 |\xi|^{2p+2} d\xi \\ &\le Ch^{1-2s+2(p+1-1/2)} \|R_h f\|^2_{H^{p+1}} \\ &\le Ch^{2(p+1-s)} \|f\|^2_{H^{p+1}}.\end{aligned}$$

The above estimates for (a), (b), (i), and (ii) lead to the result in (3.2). The estimate in (3.6) can be obtained using the Strang–Fix condition. For compactly supported $\Phi \in H^r(\mathbf{R})$ the proof in [21] is based on the Paley–Wiener theorem and the theory of entire functions. Equation (3.6) also holds when Φ is an r-regular function. In this case, the function defined by the series on the left-hand side of (3.6) is C^∞ (see Lemma 1 in [15]) and has a zero of order at least $2m+2$ at $\xi = 0$. □

This theorem can be used to derive the order of accuracy in biorthogonal multiresolution approximations.

COROLLARY 3.5. *Let P_j be the projection operators*

$$P_j f(x) = \sum_{k\in\mathbf{Z}} \langle f, \Phi^*_{jk}\rangle \Phi_{jk}(x),$$

where Φ and Φ^* are dual scaling functions associated with a biorthogonal multiresolution analysis. Assume that Φ is r-regular and $x^n\Phi^*(x) \in L^2(\mathbf{R})$ for all integer $n \ge 0$. If $f \in H^{r+1}(\mathbf{R})$ then

$$\|f - P_j f\|_{H^s} \le C 2^{-j(r+1-s)} \|f\|_{H^{r+1}},$$

for $0 \le s \le r$.

Proof. Note that $P_j f$ is written in the form (1.3)–(1.4) with $\mu(x) = \Phi^*(-x)$. As already mentioned in §2, r-regular scaling functions Φ satisfy the Strang–Fix condition with m at least equal to r. Biorthonormality implies that

$$\sum_{k\in\mathbf{Z}} \widehat{\Phi}(\xi + 2k\pi)\overline{\widehat{\Phi^*}(\xi + 2k\pi)} = 1$$

for all $\xi \in \mathbf{R}$. Furthermore, this series defines a C^∞ function (cf. Lemma 1 in [15]), and each term $\widehat{\Phi}(\xi + 2k\pi), k \neq 0$, has a zero of order $r+1$ at $\xi = 0$. Therefore

$$\widehat{\Phi}(\xi)\widehat{\mu}(\xi) = \widehat{\Phi}(\xi)\overline{\widehat{\Phi^*}(\xi)} = 1 - \sum_{k\neq 0} \widehat{\Phi}(\xi + 2k\pi)\overline{\widehat{\Phi^*}(\xi + 2k\pi)} = 1 + O(\xi^{r+1}).$$

Since the hypotheses of Theorem 3.4 are satisfied, the estimate follows. □

Remark 3.6. For orthogonal projections Π_j associated to r-regular multiresolution analysis of $L^2(\mathbf{R})$, a sharper result was obtained by Y. Meyer (cf. [18]). Let $\epsilon_j = 2^{js}\|D_j f\|_{L^2}$, where D_j is the orthogonal projection onto W_j. Recall that W_j is the orthogonal complement of V_j in V_{j+1} and

$$L^2(\mathbf{R}) = V_j \oplus_{\nu\ge j} W_\nu.$$

Then the H^s-norm is equivalent to the sum of the L^2-norm of $\Pi_0 f$ plus the ℓ^2-norm of the sequence $\epsilon_j, j \geq 0$.

When a smooth function f has a nearly singular local behavior in space or frequency domain, the H^s-norm of its derivatives becomes high, so that in practice, the order of accuracy theoretically predicted is only detected for very fine scales. Typical cases are the very oscillatory functions and the functions that undergo rapid changes in a narrow boundary layer. The approximation of such functions is particularly important in the numerical analysis of partial differencial equations. The two examples presented below are also discussed in [11] for Legendre and Chebyshev series, and also in [12] for the Zak transform. In what follows $\Phi = \Phi_N$ is the Daubechies scaling function supported on $[0, 2N-1]$, $\Psi = \Psi_N$ is the associated basic wavelet, and Π_j are the orthogonal projectors on the multiresolution analysis subspaces V_j.

Example 3.7. *Approximation of oscillatory functions.* In order to illustrate the good localization property of wavelet series in frequency domain, we will study the rate of convergence of $\Pi_j f$ in $L^2([0,1])$ for the periodic function $f(x) = \cos(2M\pi x)$.

$$\Pi_j f(x) - f(x) = \sum_{m \geq j} \sum_k b_{mk} \Psi_{mk}(x), \tag{3.7}$$

where

$$\begin{aligned} b_{mk} &= 2^{m/2} \int_{-\infty}^{\infty} \Psi(2^m x - k) \cos(2M\pi x) dx \\ &= 2^{-m/2} \int_{-\infty}^{\infty} \Psi(v) \cos(2^{-m+1} M\pi (v + k)) dv \\ &= 2^{-m/2} F_\Psi(2^{-m+1}M) \cos[2^{-m+1} Mk\pi + \eta_\Psi(2^{-m+1}M)]. \end{aligned}$$

The functions introduced above are

$$F_\Psi(\xi) = \sqrt{A_\Psi^2(\xi) + B_\Psi^2(\xi)}$$

and

$$\eta_\Psi(\xi) = \arctan \frac{B_\Psi(\xi)}{A_\Psi(\xi)},$$

where

$$A_\Psi(\xi) = \int_{-\infty}^{\infty} \Psi(v) \cos(\pi \xi v) dv,$$

$$B_\Psi(\xi) = \int_{-\infty}^{\infty} \Psi(v) \sin(\pi \xi v) dv.$$

Observe that $F_\Psi(\xi) = \sqrt{2\pi} |\widehat{\Psi}(\pi\xi)|$. Since $\widehat{\Psi}(\xi)$ has a zero of order N at $\xi = 0$ (cf.[8]), it follows that (3.7) starts converging rapidly to zero when $2^{-j+1} M\pi \sim 0$. In fact, as can be deduced from Fig. 3.1, the error $\|\Pi_j f - f\|_{L^2}$ is almost a constant, since $\Pi_j f$ is practically equal to zero for $\omega = 2^{j-1}/M < 1$ (i.e., $2^{-j+1} M\pi > \pi$). When ω increases beyond 1, $\|\Pi_j f - f\|_{L^2}$ starts decreasing, algebraically, at a rate $O(\omega^{-N})$. This rate of convergence is consistent with the estimate (3.4).

The localization of the turning point at $\omega = 1$ is related to the *bandwidth* for the filter determined by Φ. In [8] the graphs of $\Phi = \Phi_N$ and their Fourier transform $\sqrt{2\pi}|\widehat{\Phi}(\xi)|$ are plotted for different values of N. As N increases, one observes that the energy in the Fourier domain is concentrated in an interval around $\xi = 0$. Because of that, Φ can be interpreted as a *low pass filter*, in the sense that high frequencies are *attenuated* by the system $\Pi_0 f$ determined by Φ and low frequencies are *transmitted*. The function $F_\Phi(2\xi) = \sqrt{2\pi}|\widehat{\Phi}(2\pi\xi)|$ is usually called the *gain*. For a filter which has its maximum gain at ξ_0, the bandwidth is defined as $\beta = \xi_1 - \xi_2$, where ξ_1 and ξ_2 are chosen so that

$$F_\Phi^2(2\xi_1) = F_\Phi^2(2\xi_2) = F_\Phi^2(2\xi_0)/2.$$

In the present case, the maximum gain is attained at $\xi_0 = 0$ and $F_\Phi^2(0) = 2\pi$. Since $F_\Phi(2\xi)$ is a symmetric function, the bandwidth is equal to $2\xi_1$, where ξ_1 is such that

$$|\widehat{\Phi}(2\pi\xi_1)|^2 = 1/2.$$

The values of $2\pi\xi_1$ are approximately 3.115, 3.131, 3.137, 3.139, and 3.141 for $N = 3, 4, 5, 6$, and 8, respectively. One clearly sees that $2\pi\xi_1$ approaches π as N increases. That is, $\beta \sim 1$. Analogously, the bandwidth β_j corresponding to Π_j is 2^j. This implies that, as $j \to \infty$, higher frequencies are transmitted by the system Π_j. In our present case, $f(x) = \cos(2M\pi x)$ has M complete wavelengths within the interval $[0, 1]$. We argue then that the wavelet series (3.7) will start to converge rapidly to zero when $M < 2^{j-1}$, i.e., $\omega = 2^{j-1}/M > 1$. Heuristically, we conclude that the rate of convergence 2^{-jN} as given by (3.4) is only achieved when the highest frequency is transmitted.

Example 3.8. *Resolution of thin boundary layers.* Next, we will illustrate the convergence of $\Pi_j f$ for the functions

$$f_\delta(x) = \exp\left[(x-1)/\delta\right].$$

As $\delta \to 0$, $f_\delta(x)$ develops a boundary layer of width δ near $x = 1$.

It can be shown that

$$\frac{f_\delta(x) - \Pi_j f_\delta(x)}{f_\delta(x)} = E(2^j x, w),$$

where $w = 2^{-j}\delta$,

$$E(u, w) = 1 - \sum_k \Phi(u-k)\exp[-(u-k)w] \int_{-\infty}^{\infty} \Phi(v)\exp(wv)dv.$$

For fixed $w \geq 0$, $E(u, w)$ is a bounded 1-periodic function in the variable u, and for fixed $u \in \mathbf{R}$, $E(u, w)$ is a C^∞-function in the variable w. The Strang–Fix condition, together with the moment relation (3.3), implies that

$$E(u, w) = \left[\sum_{i=0}^{N} \binom{N}{i} (-1)^i \lambda(N-i)\lambda(i) + T_N(u)\right] \frac{w^N}{(N)!}$$

$$+ O(w^{N+1})$$

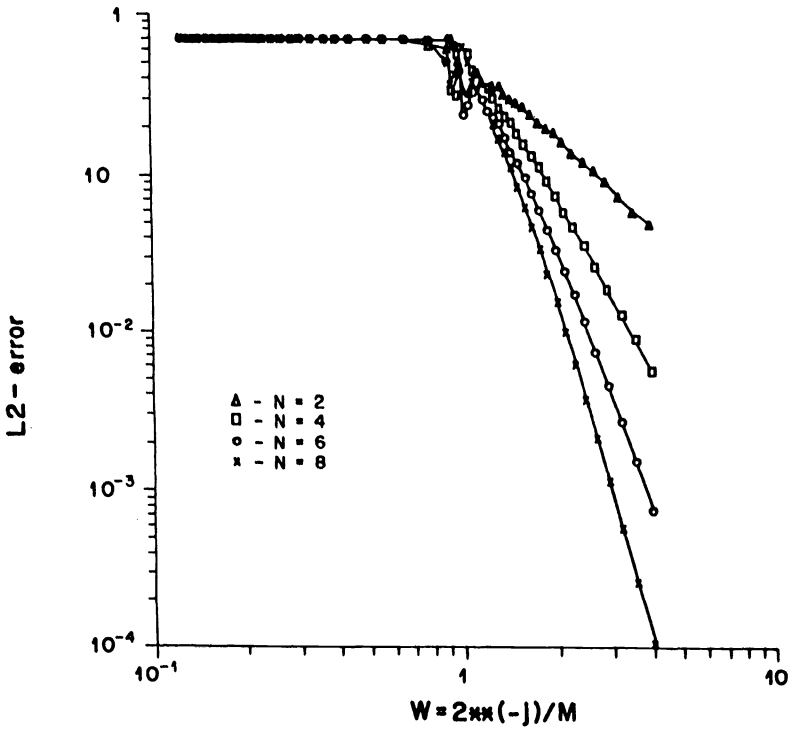


FIG. 3.1. *The L^2-error corresponding to Example* 3.7.

uniformly for $0 \leq u \leq 1$. Here T_N is the bounded and 1-periodic function defined by $T_N(x) = \sum_{k \in \mathbf{Z}} \Phi(x-k)(x-k)^N$. Therefore

$$\frac{f_\delta(x) - \Pi_j f_\delta(x)}{f_\delta(x)} = E(2^j x - [2^j x], \omega) = O(w^N),$$

as $w \to 0$. This result is consistent with the estimate (3.4) given the fact that $f_\delta(x)\delta^{-N} = \frac{d^N f_\delta}{dx^N}(x)$. It is also what one obtains using piecewise polynomials of degree $\leq N-1$ based on evenly spaced knots. It must be mentioned that better results are achieved with nonuniformly spaced knots that have higher concentration within the boundary layer, as in Chebychev expansions (cf. [11]).

REFERENCES

[1] J. P. AUBIN, *Approximation of Elliptic Boundary-value Problems*, in Pure and Applied Mathematics. A Series of Texts and Monographs, Wyley Interscience, New York, 1972.

[2] P. L. BUTZER AND R. J. NESSEL, *Fourier Analysis and Approximation Theory*, Academic Press, New York, 1971.

[3] P. L. BUTZER, S. RIES, AND R. L. STENS, *Approximations of continuous and discontinuous functions by generalized sampling series*, J. Approx. Theory, 50 (1987), pp. 23–39.

[4] P. L. BUTZER AND R. L. STENS, *Sampling series for nonnecessarily band-limited functions: An historical overview*, SIAM Rev., 34 (1992), pp. 40–53.

[5] A. COHEN, I. DAUBECHIES, AND J. C. FEAUVEAU, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[6] J. M. COMBES, A. GROSSMANN, AND P. TCHATMITCHIAN, *Wavelets, time frequency methods and phase space*, in Inverse Problems and Theoretical Imaging, Springer, Berlin, New York, 1989.

[7] W. DAHMEN AND C. A. MICCHELLI, *Translates of multivariate spline*, Linear Algebra Appl., 52/53 (1983), pp. 217–234.

[8] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[9] ———, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[10] T. EIROLA, *Sobolev characterization of solutions of dilation equations.* SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[11] D. GOTTLIEB AND S. ORZSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1977.

[12] A. J. JANSSEN, *The Zak transform: A signal transform for sampled time-continuous signals.* Philips J. Res., 43 (1988), pp. 23–69.

[13] A. KUMAR, D. R. FUHRMANN, M. FRAZIER, AND B. D. JAWERTH, *A new transform for time-frequency analysis*, IEEE Trans. Signal Processing, 40 (1992), pp. 1697–1707.

[14] P. G. LEMARIÉ, *Les ondelettes en 1989*, in Lecture Notes in Mathematics 1438, Springer-Verlag, Berlin, 1990.

[15] ———, *Functions a support compact dans les analyses multirésolutions*, Rev. Mat. Iberoamericana, 7 (1991), pp. 157–182.

[16] ———, *Analyses multirésolutions non orthogonales, commutation entre prokecteurs et derivation et ondelettes vecteurs à divergence nulle*, Rev. Mat. Iberoamericana, 8 (1992), pp. 221–237.

[17] S. MALLAT, *Multiresolution approximations and wavelet orthonormal basis of* $L^2(\mathbf{R})$, Trans. Amer. Math. Soc., 315 (1991), pp. 344–351.

[18] Y. MEYER, *Ondelettes et Operateurs*, Hermann, Paris, 1990.

[19] ———, *Wavelets and applications*, in Proceedings of the International Conference, Marseille, France, May 1990. Y. Meyer, ed., Springer-Verlag, Berlin, 1992.

[20] G. RICHARDS, *A Gibbs phenomenon for spline functions*, J. Approx. Theory, 66 (1991), pp. 344–351.

[21] G. STRANG AND G. A. FIX, *A Fourier analysis of the finite element method*, in Constructive Aspects of Functional Analysis, Edizioni Cremonese, Rome, 1973.

[22] M. VETTERLI AND C. HERLEY, *Wavelets and filter banks: Theory and design*, IEEE Trans. Signal Processing, 40 (1992), pp. 2207–2232.

[23] G. WALTER, *A sampling theorem for wavelet subspaces*, IEEE Trans. Inform. Theory, 38 (1992), pp. 881–884.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1403-1414, November 1995

STABILITY FOR SYSTEMS OF CONSERVATION LAWS IN SEVERAL SPACE DIMENSIONS*

C. M. DAFERMOS†

Abstract. It is shown that admissible L^∞ solutions of any system of two hyperbolic conservation laws in several space variables with flux functions whose Jacobians commute are always stable in L^p for $p \in [1,2]$ and, under additional assumptions, also stable in L^∞.

Key words. hyperbolic conservation laws, entropy, invariant regions

AMS subject classification. 35L65

1. Introduction. We consider the initial-value problem for a hyperbolic system of n conservation laws in m space variables:

$$\partial_t U(x,t) + \sum_{\alpha=1}^{m} \partial_\alpha F^\alpha(U(x,t)) = 0, \quad x \in \mathbb{R}^m, \ t \in (0,\infty), \tag{1.1}$$

$$U(x,0) = U_0(x), \quad x \in \mathbb{R}^m. \tag{1.2}$$

The symbol ∂_t stands for $\partial/\partial t$ and ∂_α denotes $\partial/\partial x^\alpha$, $\alpha = 1,\dots,m$. The *state vector* U takes values in an open bounded neighborhood $\mathcal{O}$ of the origin in $\mathbb{R}^n$. For $\alpha = 1,\dots,m$, F^α are given smooth maps from $\mathcal{O}$ to $\mathbb{R}^n$ such that for each unit vector ν in $\mathbb{R}^m$ and any fixed U in $\mathcal{O}$, the $n \times n$ matrix

$$\sum_{\alpha=1}^{m} \nu_\alpha DF^\alpha(U) \tag{1.3}$$

has real eigenvalues and n linearly independent eigenvectors.

When the system is nonlinear, the initial-value problem is notoriously difficult. Solutions starting out from smooth initial values eventually develop discontinuities; hence only a theory of weak solutions is relevant, globally in time. Accordingly, here we prescribe initial data U_0 in $L^\infty(\mathbb{R}^m;\mathcal{O})$ and consider weak solutions U in $L^\infty(\mathbb{R}^m \times (0,\infty);\mathcal{O})$. The reader should be aware, however, that the existence of such solutions has been established in a definitive manner only when $n = 1$ [7].

To secure uniqueness within the class of weak solutions, one has to adopt *admissibility criteria* that will disqualify spurious solutions (for a survey, see [4]). Here we will enforce admissibility by requiring solutions to satisfy "entropy" inequalities of the form

$$\partial_t \eta(U(x,t)) + \sum_{\alpha=1}^{m} \partial_\alpha q^\alpha(U(x,t)) \le 0, \tag{1.4}$$

where η, the *entropy*, is a smooth real-valued function on $\mathcal{O}$, and $q = (q^1,\dots,q^m)$, the associated *entropy flux*, is a smooth map from $\mathcal{O}$ to $\mathbb{R}^m$. We postulate that every Lipschitz continuous solution of (1.1) is admissible and thus satisfies (1.4) automatically.

* Received by the editors November 12, 1993; accepted for publication March 4, 1994. This research was supported in part by National Science Foundation grant DMS-9208284, Army Research Office contract DAAH04-93-G-0125, and Office of Naval Research contract N00014-92-J-1481.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

This will be the case if and only if

$$Dq^\alpha = D\eta DF^\alpha, \quad \alpha = 1, \dots, m. \tag{1.5}$$

We require that every admissible solution of (1.1) satisfy (1.4) for any pair (η, q), with η convex, for which (1.5) holds. The compatibility conditions

$$H\eta DF^\alpha = (DF^\alpha)^T H\eta, \quad \alpha = 1, \dots, m \tag{1.6}$$

for (1.5), where $H\eta$ denotes the Hessian matrix of η, generally induce $\frac{1}{2}n(n-1)m$ independent equations on η. Therefore, unless either $n = 1$ and m is arbitrary or $n = 2$ and $m = 1$, system (1.1) should not be expected to possess entropy–entropy flux pairs beyond the trivial ones: $\eta(U) = P^T U$, $q^\alpha(U) = P^T F^\alpha(U)$ with P some constant n-vector. Nevertheless, the systems of conservation laws of continuum physics are endowed with a nontrivial entropy–entropy flux pair for which (1.4) expresses, explicitly or implicitly, the second law of thermodynamics. This entropy, however, is not necessarily convex (see [5]).

When (1.1) possesses an entropy–entropy flux pair (η, q) with η convex and

$$|q(U)| \le s\eta(U), \quad U \in \mathcal{O}, \tag{1.7}$$

then, as shown in the appendix,

$$\int_{|x|<r} \eta(U(x,t))dx \le \int_{|x|<r+st} \eta(U_0(x))dx, \quad 0 < t < \infty \tag{1.8}$$

for any $r > 0$. In particular, when $\eta(U)$ is uniformly convex,[1]

$$\int_{|x|<r} |U(x,t)|^2 dx \le c_2 \int_{|x|<r+st} |U_0(x)|^2 dx, \quad 0 < t < \infty. \tag{1.9}$$

The natural question of whether admissible solutions of (1.1), (1.2) also satisfy stability estimates

$$\int_{|x|<r} |U(x,t)|^p dx \le c_p \int_{|x|<r+st} |U_0(x)|^p dx, \quad 0 < t < \infty \tag{1.10}$$

for $p \neq 2$ then arises. The cases $p = 1$ and $p \to \infty$ are of particular interest: the former would hint that (1.1), (1.2) may be well posed in BV, while the latter would induce valuable pointwise bounds on solutions. Estimates (1.10) with $c_p = 1$, for any $p \in [1, \infty)$, are well known for the single equation $n = 1$.

As shown by Brenner [1], when (1.1) is linear, (1.10) holds for all $p \in [1, \infty)$ if and only if the Jacobians of the F^α commute:

$$DF^\alpha DF^\beta = DF^\beta DF^\alpha, \quad \alpha, \beta = 1, \dots, m. \tag{1.11}$$

Rauch [9] notes that if solutions of a quasilinear system (1.1) satisfy (1.10) then the same must be true for solutions of the system resulting from linearization about any constant state. Therefore, (1.11) is a necessary condition for (1.10), with $p \neq 2$, in

[1] As noted by Friedrichs and Lax [6], when η is uniformly convex, the change of variables $V = D\eta(U)$ renders (1.1) symmetric. Conversely, if (1.1) is symmetric then $\eta(U) = \frac{1}{2}|U|^2$ is a uniformly convex entropy with entropy flux $q, q^\alpha = \int U^T DF^\alpha(U)dU$.

the quasilinear case as well. The object of this paper is to demonstrate that (1.11) is also a sufficient condition for (1.10), at least when $n = 2$.

Our approach will be based on the following observation. Since (1.1) is assumed to be hyperbolic, for $\alpha = 1, \dots, m$, $DF^\alpha(U)$ will have real eigenvalues and n linearly independent eigenvectors. When (1.11) holds, all the $DF^\alpha(U)$ share the same set of n linearly independent eigenvectors, say $R_1(U), \dots, R_n(U)$. It is easily checked that, upon renormalizing, if necessary, the eigenvectors associated with multiple eigenvalues, (1.6) is equivalent to

$$R_i^T H\eta R_j = 0, \quad i, j = 1, \dots, n, \quad i \neq j \tag{1.12}$$

(see [3] also). Thus, under assumption (1.11) the number of independent conditions on η is reduced from $\frac{1}{2}n(n-1)m$ to $\frac{1}{2}n(n-1)$. In particular, when $n = 2$, (1.12) reduces to a single linear hyperbolic differential equation of the second order, which admits a large family of solutions. Consequently, we will limit our investigation to the case $n = 2$. The analysis may be readily extended to the class of systems of more than two equations, identified by Serre [10], which are endowed with a "rich" family of entropies.

To establish (1.10) it would suffice to construct a convex solution η of (1.12) such that

$$c|U|^p \leq \eta(U) \leq C|U|^p, \quad U \in \mathcal{O} \tag{1.13}$$

for positive constants c, C, and associated entropy flux q satisfying (1.7). Note that, because of (1.12), η will be convex if and only if

$$R_i^T H\eta R_i \geq 0, \quad i = 1, \dots, n. \tag{1.14}$$

For $n = 2$ and any $p \in [1, 2]$, in §3 we construct entropy–entropy flux pairs (η, q) with the above specifications thus establishing L^p stability (1.10) for p in that range. To attain L^p stability when $p > 2$, special restrictions have to be imposed on system (1.1). Conditions that induce L^p stability for $p >> 1$ and that, in the limit $p \to \infty$, yield positively invariant regions for solutions are discussed in §3.

It should be noted that the aforementioned conclusions on L^p stability and invariant regions for the hyperbolic system (1.1) apply equally well to the parabolic system

$$\partial_t U + \sum_{\alpha=1}^{m} \partial_\alpha F^\alpha(U) = \kappa \Delta U, \tag{1.15}$$

because any solution of it satisfies the identity

$$\partial_t \eta + \sum_{\alpha=1}^{m} \partial_\alpha q^\alpha = \kappa \Delta \eta - \kappa \sum_{\alpha=1}^{m} \partial_\alpha U^T H\eta \partial_\alpha U \tag{1.16}$$

for any entropy–entropy flux pair (η, q) of (1.1). The relation between invariant regions obtained through entropy inequalities and those deduced from the maximum principle will be discussed in §3.

Clearly, the class of systems in which DF^α and DF^β commute is very special. In §2 we shall see how one may construct systems of two conservation laws with this property, starting out from assigned eigenvectors. The simple example

$$F^\alpha(U) = e_\alpha(|U|^2)U, \quad \alpha = 1, \dots, m, \tag{1.17}$$

where e_α, $\alpha = 1, \dots, m$, are smooth real-valued functions, may provide a useful model for exploring other special properties that systems with commuting Jacobians potentially have (simpler solution of the Riemann problem, BV estimates, etc.).

2. Riemann invariants and entropies. We consider a hyperbolic system of two conservation laws (1.1) with the property (1.11). Thus, for $\alpha = 1, \dots, m$ and any $U \in \mathcal{O}$, the matrix $DF^\alpha(U)$ has real eigenvalues $\lambda_\alpha(U) \leq \mu_\alpha(U)$, generally depending on α, and a corresponding set of linearly independent eigenvectors $R(U)$, $S(U)$ independent of α.

Upon renormalizing, if necessary, R and S, we construct real-valued fields z and w on $\mathcal{O}$ such that

$$\begin{cases} DzR = 1, & DzS = 0, \\ DwR = 0, & DwS = 1. \end{cases} \tag{2.1}$$

In analogy with the one-space dimensional situation, we will call z and w *Riemann invariants* of the system. We note that for $\alpha = 1, \dots, m$ and $U \in \mathcal{O}$, $Dz(U)$ and $Dw(U)$ are left (row) eigenvectors of $DF^\alpha(U)$ with associated eigenvalues $\lambda_\alpha(U)$ and $\mu_\alpha(U)$, respectively. In particular,

$$DF^\alpha = \lambda_\alpha R Dz + \mu_\alpha S Dw. \tag{2.2}$$

If U is a classical solution of (1.1), multiplying the system from the left by Dz or Dw we obtain

$$\begin{cases} \partial_t z + \sum_{\alpha=1}^m \lambda_\alpha \partial_\alpha z = 0, \\ \partial_t w + \sum_{\alpha=1}^m \mu_\alpha \partial_\alpha w = 0, \end{cases} \tag{2.3}$$

which shows that z stays constant along 1-characteristics and w stays constant along 2-characteristics.

For simplicity let us assume that the map $U \mapsto (z, w)$ is a diffeomorphism of $\mathcal{O}$ to a rectangle $\mathcal{R}$ of $\mathbb{R}^2$, mapping 0 to 0. Therefore, we may use new coordinates (z, w) in place of the original state vector U. To avoid cumbersome notation, we shall be employing, in the customary fashion, the same symbol to express any particular field as a function of U and as a function of (z, w). By virtue of (2.1), for the typical field, say ρ, the chain rule yields

$$\rho_z = D\rho R, \quad \rho_w = D\rho S. \tag{2.4}$$

We show that $\lambda_\alpha, \mu_\alpha$ are fields of eigenvalues of the Jacobian DF^α of a vector field F^α with associated eigenvector fields R, S if and only if

$$(\lambda_\alpha R)_w - (\mu_\alpha S)_z = 0. \tag{2.5}$$

Indeed, let z, w be determined through (2.1) and denote by A^α the 2×2 matrix field on the right-hand side of (2.2), which has eigenvalues $\lambda_\alpha, \mu_\alpha$, right (column) eigenvectors R, S, and left (row) eigenvectors Dz, Dw. We note that $A^\alpha R_w - A^\alpha S_z = 0$, because

$$\begin{cases} Dz[A^\alpha R_w - A^\alpha S_z] = \lambda_\alpha [DzR_w - DzS_z] = -\lambda_\alpha [(Dz)_w R - (Dz)_z S] = 0, \\ Dw[A^\alpha R_w - A^\alpha S_z] = \mu_\alpha [DwR_w - DwS_z] = -\mu_\alpha [(Dw)_w R - (Dw)_z S] = 0. \end{cases} \tag{2.6}$$

Therefore

$$(\lambda_\alpha R)_w - (\mu_\alpha S)_z = (A^\alpha R)_w - (A^\alpha S)_z = A^\alpha_w R - A^\alpha_z S. \tag{2.7}$$

By virtue of (2.4) the right-hand side of (2.7) vanishes if and only if DA^α is symmetric, i.e., if and only if $A^\alpha = DF^\alpha$ for some vector field F^α.

Consequently, the entire class of systems (1.1) of two conservation laws with the property (1.11) may be constructed by the following procedure. We start out from arbitrarily assigned linearly independent fields R, S, determine z, w through (2.1), then select any solutions $\lambda_\alpha, \mu_\alpha$ of (2.5), $\alpha = 1, \dots, m$, and finally get $F^\alpha, \alpha = 1, \dots, m$, from (2.2).

By virtue of (1.12), (1.14), to construct convex entropies η of (1.1) in the present setting we have to find solutions of the equation

$$R^T H\eta S = 0 \tag{2.8}$$

that satisfy the inequalities

$$R^T H\eta R \geq 0\,, \quad S^T H\eta S \geq 0. \tag{2.9}$$

Because of (2.4),

$$\begin{aligned} R^T H\eta S &= D(D\eta R)S - D\eta DRS \\ &= \eta_{zw} - \eta_z DzR_w - \eta_w DwR_w, \end{aligned} \tag{2.10}$$

$$\begin{aligned} R^T H\eta R &= D(D\eta R)R - D\eta DRR \\ &= \eta_{zz} - \eta_z DzR_z - \eta_w DwR_z, \end{aligned} \tag{2.11}$$

$$\begin{aligned} S^T H\eta S &= D(D\eta S)S - D\eta DSS \\ &= \eta_{ww} - \eta_z DzS_w - \eta_w DwS_w. \end{aligned} \tag{2.12}$$

On the other hand, using (2.1) and (2.4),

$$\begin{cases} -DzR_w = (Dz)_w R = S^T HzR := a, \\ -DwR_w = (Dw)_w R = S^T HwR := b, \end{cases} \tag{2.13}$$

$$\begin{cases} -DzR_z = (Dz)_z R = R^T HzR := f, \\ -DwR_z = (Dw)_z R = R^T HwR := g, \end{cases} \tag{2.14}$$

$$\begin{cases} -DzS_w = (Dz)_w S = S^T HzS := h, \\ -DwS_w = (Dw)_w S = S^T HwS := k, \end{cases} \tag{2.15}$$

where Hz and Hw denote the Hessian matrices of z and w. Thus (2.8) and (2.9) reduce to

$$\eta_{zw} + a\eta_z + b\eta_w = 0, \tag{2.16}$$

$$\begin{cases} \eta_{zz} + f\eta_z + g\eta_w \geq 0, \\ \eta_{ww} + h\eta_z + k\eta_w \geq 0. \end{cases} \tag{2.17}$$

We note that the coefficients a, b, f, g, h, k appearing in (2.16), (2.17) and defined through (2.13), (2.14), (2.15) arise in several connections in the theory of hyperbolic

systems of two conservation laws. For example, the sign of h (or g) determines the direction in which z (or w) jumps across admissible 2-shocks (or 1-shocks).

To determine the entropy flux q associated with an entropy η, we multiply (1.5) from the right by R or S and use (2.4) to get

$$q_z^\alpha = \lambda_\alpha \eta_z, \quad q_w^\alpha = \mu_\alpha \eta_w, \quad \alpha = 1, \dots, m. \tag{2.18}$$

Eliminating q^α between the two equations in (2.18) (assuming $\lambda_\alpha \neq \mu_\alpha$), yields the familiar equation

$$\eta_{zw} + \frac{\lambda_{\alpha w}}{\lambda_\alpha - \mu_\alpha}\eta_z + \frac{\mu_{\alpha z}}{\mu_\alpha - \lambda_\alpha}\eta_w = 0, \quad \alpha = 1, \dots, m, \tag{2.19}$$

which may leave the impression that η depends on the characteristic speeds. However, multiplying (2.5) from the left by Dz or by Dw and using (2.1), (2.4), and (2.13), we deduce

$$\frac{\lambda_{\alpha w}}{\lambda_\alpha - \mu_\alpha} = a, \quad \frac{\mu_{\alpha z}}{\mu_\alpha - \lambda_\alpha} = b, \quad \alpha = 1, \dots, m, \tag{2.20}$$

which shows that (2.19) is identical to (2.16).

3. Stability in L^p. The aim here is to construct entropy–entropy flux pairs for (1.1) inducing the L^p stability property (1.10) for various values of p. As before, we focus our investigation on systems of two conservation laws, $n = 2$, satisfying (1.11). The Riemann invariants (z, w) will take values in a rectangle $\mathcal{R}$, which is assumed small so that terms of order $O(|z|^\gamma + |w|^\gamma), \gamma > 0$, will be small in comparison to 1. However, the upper bounds on the size of $\mathcal{R}$ and the constants involved in the order symbols which appear throughout this section will be uniform with respect to $p \in (1, \infty)$, so we will have no difficulty in treating the limiting cases $p = 1$ and $p = \infty$.

LEMMA 3.1. *For any p in $(1, \infty)$ there is an entropy-entropy flux pair (η, q) with the following properties:*

$$\eta(z, w) = |z|^p + |w|^p + O(|z| + |w|)(|z|^p + |w|^p), \tag{3.1}$$

$$\begin{aligned} q^\alpha(z, w) &= \lambda_\alpha(0,0)|z|^p + \mu_\alpha(0,0)|w|^p + O(|z| + |w|)(|z|^p + |w|^p), \\ &\alpha = 1, \dots, m, \end{aligned} \tag{3.2}$$

$$\begin{aligned} \eta_{zz} + f\eta_z + g\eta_w &= p(p-1)[1 + O(|w|)]|z|^{p-2} + pO(|z|^{p-1}) \\ &+ p[1 + O(|z|)]\{g(0,0)|w|^{p-2}w + g_z(0,0)|w|^{p-2}wz + g_w(0,0)|w|^p\} \\ &+ [b^2(0,0) - b(0,0)f(0,0) - b_z(0,0)]|w|^p + pO(|z| + |w|)(|z|^p + |w|^p), \end{aligned} \tag{3.3}$$

$$\begin{aligned} \eta_{ww} + h\eta_z + k\eta_w &= p(p-1)[1 + O(|z|)]|w|^{p-2} + pO(|w|^{p-1}) \\ &+ p[1 + O(|w|)]\{h(0,0)|z|^{p-2}z + h_w(0,0)|z|^{p-2}zw + h_z(0,0)|z|^p\} \\ &+ [a^2(0,0) - a(0,0)k(0,0) - a_w(0,0)]|z|^p + pO(|z| + |w|)(|z|^p + |w|^p). \end{aligned} \tag{3.4}$$

Proof. Let $\eta(z, w)$ be the solution of (2.16) on $\mathcal{R}$ under Goursat-type conditions

$$\eta(z, 0) = |z|^p, \quad \eta(0, w) = |w|^p. \tag{3.5}$$

We fix any $(z,w) \in \mathcal{R}$, integrate (2.16) over the rectangle with vertices $(0,0)$, $(z,0)$, (z,w), $(0,w)$ and perform integrations by parts to get

$$
\begin{aligned}
\eta(z,w) = {} & \eta(z,0) + \eta(0,w) \\
& - \int_0^w \{a(z,\omega)\eta(z,\omega) - a(0,\omega)\eta(0,\omega)\}d\omega \\
& - \int_0^z \{b(\zeta,w)\eta(\zeta,w) - b(\zeta,0)\eta(\zeta,0)\}d\zeta \\
& + \int_0^w \int_0^z \{a_z(\zeta,\omega) + b_w(\zeta,\omega)\}\eta(\zeta,\omega)d\zeta d\omega.
\end{aligned} \tag{3.6}
$$

From (3.5) and (3.6) we deduce (3.1) through iteration.

The associated entropy flux q may be computed from η by integrating (2.18):

$$
\begin{aligned}
q^\alpha(z,w) = {} & \mu_\alpha(z,w)\eta(z,w) + [\lambda_\alpha(z,0) - \mu_\alpha(z,0)]\eta(z,0) \\
& - \int_0^z \lambda_{\alpha z}(\zeta,0)\eta(\zeta,0)d\zeta - \int_0^w \mu_{\alpha w}(z,\omega)\eta(z,\omega)d\omega,
\end{aligned} \tag{3.7}
$$

from which (3.2) follows by virtue of (3.1).

We now multiply (2.16) by the integrating factor $\exp \int a dw$ and integrate with respect to w. After an integration by parts,

$$
\begin{aligned}
\eta_z(z,w) = {} & \exp[-\int_0^w a(z,\xi)d\xi]\eta_z(z,0) - b(z,w)\eta(z,w) \\
& + b(z,0)\exp[-\int_0^w a(z,\xi)d\xi]\eta(z,0) \\
& + \int_0^w \exp[-\int_\omega^w a(z,\xi)d\xi]\{b_w(z,\omega) + a(z,\omega)b(z,\omega)\}\eta(z,\omega)d\omega.
\end{aligned} \tag{3.8}
$$

Combining (3.8) with (3.1) and (3.5), we obtain

$$
\eta_z(z,w) = p[1 + O(|w|)]|z|^{p-2}z - b(0,0)|w|^p + O(|z| + |w|)(|z|^p + |w|^p). \tag{3.9}
$$

Similarly,

$$
\eta_w(z,w) = p[1 + O(|z|)]|w|^{p-2}w - a(0,0)|z|^p + O(|z| + |w|)(|z|^p + |w|^p). \tag{3.10}
$$

Next we differentiate (2.16) with respect to z and use (2.16) again to get

$$
\eta_{zzw} + a\eta_{zz} = (ab - a_z)\eta_z + (b^2 - b_z)\eta_w. \tag{3.11}
$$

We multiply (3.11) by the integrating factor $\exp \int a dw$ and then integrate with respect to w to obtain

(3.12)

$$
\begin{aligned}
\eta_{zz}(z,w) = {} & \exp[-\int_0^w a(z,\xi)d\xi]\eta_{zz}(z,0) + [b^2(z,w) - b_z(z,w)]\eta(z,w) \\
& - \exp[-\int_0^w a(z,\xi)d\xi][b^2(z,0) - b_z(z,0)]\eta(z,0) \\
& + \int_0^w \exp[-\int_\omega^w a(z,\xi)d\xi][a(z,\omega)b(z,\omega) - a_z(z,\omega)]\eta_z(z,\omega)d\omega \\
& - \int_0^w \exp[-\int_\omega^w a(z,\xi)d\xi][a(z,\omega)b^2(z,\omega) - a(z,\omega)b_z(z,\omega) \\
& \qquad + 2b(z,\omega)b_w(z,\omega) - b_{zw}(z,\omega)]\eta(z,\omega)d\omega.
\end{aligned}
$$

From (3.12), (3.1), (3.9), and (3.5) it follows that

$$
\begin{aligned}
\eta_{zz}(z,w) = {} & p(p-1)[1 + O(|w|)]|z|^{p-2} + pO(|w|)|z|^{p-1} \\
& + [b^2(0,0) - b_z(0,0)]|w|^p + O(|z| + |w|)(|z|^p + |w|^p).
\end{aligned} \tag{3.13}
$$

Combining (3.9), (3.10), and (3.13) we arrive at (3.3). Interchanging the roles of z and w yields (3.4). This completes the proof.

We have now laid the preparation for addressing the issue of L^p stability. Recall that the strategy is to get (1.10) via the entropy inequality (1.8) for an entropy–entropy flux pair (η, q) with η convex which satisfies (1.7) and (1.13).

Let $(\eta_p, q_p), p \in (1, \infty)$, denote the entropy–entropy flux pair constructed in Lemma 3.1. Note that (η_p, q_p) satisfies (1.7) and (1.13) for any $p \in (1, \infty)$ by virtue of (3.1) and (3.2). We shall check whether η_p is a convex function of U through (2.17) with the help of (3.3), (3.4). For $p \in (1, 2]$, the right-hand side of (3.3) is $p(p-1)|z|^{p-2} + O(|z|^{p-1} + |w|^{p-1})$ and the right-hand side of (3.4) is $p(p-1)|w|^{p-2} + O(|z|^{p-1} + |w|^{p-1})$. Therefore, for $p \in (1, 2]$, setting

$$\eta = \eta_p + \sigma\eta_2, \quad q = q_p + \sigma q_2, \tag{3.14}$$

where σ is a sufficiently large positive constant, yields an entropy–entropy flux pair (η, q) with $\eta(U)$ convex, satisfying (1.7), (1.13). Furthermore, σ may be chosen independently of $p \in (1, 2]$ so we can pass to the limit $p \downarrow 1$ in (3.14) and get an entropy–entropy flux pair (η, q) with $\eta(U)$ convex which satisfies (1.7), (1.13) for $p = 1$. We have thus established the following theorem.

THEOREM 3.1. *Consider a hyperbolic system* (1.1) *of two conservation laws with commuting Jacobians* (1.11). *Admissible solutions with small oscillation of the initial-value problem* (1.1), (1.2) *have the L^p stability property* (1.10) *for any $p \in [1, 2]$.*

We now turn to the case $p > 2$. It is no longer possible to add, as in (3.14), $\sigma\eta_2$ to η_p without violating (1.13), and so we will have to work with (η_p, q_p) as our entropy–entropy flux pair, requiring that $\eta_p(U)$ itself be convex. It follows from (3.3) and (3.4) that

$$g(0,0) = 0, \quad h(0,0) = 0 \tag{3.15}$$

are necessary conditions for (2.17) to hold. It is easy to derive sufficient conditions that would apply to particular ranges of values of p. Our goal here is to pass eventually to the limit $p \to \infty$, and so we seek sufficient conditions effective for $p >> 2$. From (3.3), (3.4) and by virtue of Young's inequality

$$\begin{cases} \left||w|^{p-2}wz\right| \le \frac{1}{p}|z|^p + \frac{p-1}{p}|w|^p, \\ \left||z|^{p-2}zw\right| \le \frac{1}{p}|w|^p + \frac{p-1}{p}|z|^p, \end{cases} \tag{3.16}$$

it is clear that (2.17) will hold for p large when, in addition to (3.15),

$$\begin{cases} g_w(0,0) - |g_z(0,0)| > 0, \\ h_z(0,0) - |h_w(0,0)| > 0. \end{cases} \tag{3.17}$$

We have thus proved the following theorem.

THEOREM 3.2. *Consider a hyperbolic system* (1.1) *of two conservation laws with commuting Jacobians* (1.11). *When* (3.15) *and* (3.17) *hold, admissible solutions with small oscillation of the initial-value problem* (1.1), (1.2) *have the L^p stability property* (1.10) *for any $p >> 2$.*

Assuming (3.15), (3.17) and $p >> 2$, we write (1.8) for the entropy η_p, raise the resulting inequality to the power $1/p$, and let $p \to \infty$. By virtue of (3.1), this yields the following theorem.

THEOREM 3.3. *Consider a hyperbolic system* (1.1) *of two conservation laws with commuting Jacobians* (1.11). *When* (3.15) *and* (3.17) *hold, the Riemann invariants of*

admissible solutions with small oscillation have the L^∞ stability property

$$\begin{aligned}&\max\{\|z(\cdot,t)\|_{L^\infty(|x|<r)},\ \|w(\cdot,t)\|_{L^\infty(|x|<r)}\}\\&\le \max\{\|z(\cdot,0)\|_{L^\infty(|x|<r+st)},\ \|w(\cdot,0)\|_{L^\infty(|x|<r+st)}\}\end{aligned}\tag{3.18}$$

for some $s>0$ and any $r>0, t>0$. In particular, for any δ positive small, the square $\{(z,w): |z|\le\delta, |w|\le\delta\}$ is a positively invariant region for the Riemann invariants of solutions.

An alternative way to show that the squares are positively invariant regions is by employing Lax-type entropy–entropy flux pairs [8]

$$\begin{cases}\eta(z,w)=\cosh(\ell z)[u(z,w)+O(\frac{1}{\ell})],\\ q^\alpha(z,w)=\cosh(\ell z)[\lambda_\alpha(z,w)u(z,w)+O(\frac{1}{\ell})],\ \alpha=1,\dots,m,\end{cases}\tag{3.19}$$

$$\begin{cases}\eta(z,w)=\cosh(\ell w)[v(z,w)+O(\frac{1}{\ell})],\\ q^\alpha(z,w)=\cosh(\ell w)[\mu_\alpha(z,w)v(z,w)+O(\frac{1}{\ell})],\ \alpha=1,\dots,m.\end{cases}\tag{3.20}$$

To satisfy (2.18) and because of (2.20), we need

$$u_w+au=0,\quad v_z+bv=0.\tag{3.21}$$

To secure that the convexity conditions (2.17) will hold as $\ell\to\infty$, we would need $u>0, v>0$, which can be easily attained through (3.21), and also $h(0,w)=0, h_z(0,w)>0, g(z,0)=0, g_w(z,0)>0$, i.e., conditions slightly stronger than (3.15), (3.17).

We note that the assertions of Theorems 3.1, 3.2, and 3.3 for the hyperbolic system (1.1) also apply to the parabolic system (1.15) with $r=\infty$ and all other assumptions remaining the same. The reason, of course, is that solutions of (1.15) satisfy the identity (1.16) for any entropy–entropy flux pair (η,q) of (1.1). When η is convex, integrating (1.16) over $\mathbb{R}^m\times[0,t]$ yields (1.8) with $r=\infty$.

As shown by Chueh, Conley, and Smoller [2], the square $\{(z,w): |z|\le\delta, |w|\le\delta\}$ will be an invariant region for (1.15) if and only if

$$\begin{cases}g(z,-\delta)\le 0,\ g(z,\delta)\ge 0,\ -\delta\le z\le\delta,\\ h(-\delta,w)\le 0,\ h(\delta,w)\ge 0,\ -\delta\le w\le\delta.\end{cases}\tag{3.22}$$

We sketch a direct proof to gain another perspective for the implications of (1.11). Upon multiplying (1.15) by Dz or by Dw from the left, we get

$$\begin{cases}\partial_t z+\sum_{\alpha=1}^m\lambda_\alpha\partial_\alpha z=\kappa\Delta z-\kappa\sum_{\alpha=1}^m\partial_\alpha U^T Hz\partial_\alpha U,\\ \partial_t w+\sum_{\alpha=1}^m\mu_\alpha\partial_\alpha w=\kappa\Delta w-\kappa\sum_{\alpha=1}^m\partial_\alpha U^T Hw\partial_\alpha U.\end{cases}\tag{3.23}$$

Now suppose that some solution of (1.15), whose range was initially confined inside the square $\{(z,w): |z|<\delta, |w|<\delta\}$, crosses the boundary of the square for the first time at time $\bar t>0$ at a point $\bar x$. For definiteness, assume $z(\bar x,\bar t)=\delta$ (the other possibilities can be handled by the same method). Thus at $(\bar x,\bar t)$, $\partial_\alpha z=0$, i.e., $Dz\partial_\alpha U=0$, and so $\partial_\alpha U$ is collinear to S. Then, by (2.15) and (3.22), $\sum_{\alpha=1}^m\partial_\alpha U^T Hz\partial_\alpha U\ge 0$ at $(\bar x,\bar t)$. The classical maximum principle for parabolic equations now provides the desired contradiction.

Note that (3.15) and (3.17) imply that (3.22) holds for any positive small δ. Serre [11] has developed a method based on entropy inequalities, which could be used to show that, for the square $\{(z,w): |z|\le\delta, |w|\le\delta\}$ to be an invariant region for (1.1), it would suffice to assume (3.22) just for that particular δ.

Appendix. Here we verify the assertion that if (η, q) is an entropy–entropy flux pair for a system (1.1) of conservation laws such that η is convex and (1.7) holds, then any L^∞ solution of (1.1), (1.2) which conforms with (1.4) satisfies (1.8) for every $r > 0, t > 0$. Of course, when U is of class BV, (1.8) is simply established by integrating (1.4) over the frustum $\{(x,\tau) : 0 < \tau < t,\ |x| < r + s(t-\tau)\}$ and applying Green's theorem.

We begin by showing that after redefining, if necessary, U on a set of measure zero,

$$\int_{\mathbb{R}^m} \chi(x)U(x,t)dx = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \int_{t-\varepsilon}^{t} \int_{\mathbb{R}^m} \chi(x)U(x,\tau)dxd\tau,\ \ 0 < t < \infty \tag{A.1}$$

holds for any $\chi \in L^1(\mathbb{R}^m)$. Let $\chi \in C_0^\infty(\mathbb{R}^m)$. We fix $t > 0, \varepsilon > 0$ and construct the Lipschitz function ϕ, with compact support in $\mathbb{R}^m \times [0,\infty)$, by $\phi(x,\tau) = \chi(x)\theta(\tau)$, where

$$\theta(\tau) = \begin{cases} 1, & 0 \le \tau < t - \varepsilon, \\ \frac{1}{\varepsilon}(t-\tau), & t - \varepsilon \le \tau < t, \\ 0, & t \le \tau < \infty. \end{cases} \tag{A.2}$$

Then

$$\begin{aligned} 0 &= \int_0^\infty \int_{\mathbb{R}^m} \left\{ -\partial_t \phi U - \sum_{\alpha=1}^m \partial_\alpha \phi F^\alpha(U) \right\} dxd\tau - \int_{\mathbb{R}^m} \phi(x,0)U_0(x)dx \\ &= \frac{1}{\varepsilon} \int_{t-\varepsilon}^{t} \int_{\mathbb{R}^m} \chi(x)U(x,\tau)dxd\tau - \int_0^t \int_{\mathbb{R}^m} \sum_{\alpha=1}^m \partial_\alpha \chi(x) F^\alpha(U(x,\tau))dxd\tau \\ &\quad - \int_{\mathbb{R}^m} \chi(x)U_0(x)dx + O(\varepsilon). \end{aligned} \tag{A.3}$$

It follows that for any $t > 0$, the limit on the right-hand side of (A.1) exists for all $\chi \in C_0^\infty(\mathbb{R}^m)$ and, thereby, for all $\chi \in L^1(\mathbb{R}^m)$. We may thus normalize U so that (A.1) holds. This renders the map $t \mapsto U(\cdot, t)$ continuous in L^∞ weak *.

Next we fix $r > 0, t > 0, \varepsilon > 0$ and construct the Lipschitz function ϕ by $\phi(x,\tau) = \psi(x,\tau)\theta(\tau)$, where θ is again defined by (A.2) and

$$\psi(x,\tau) = \begin{cases} 1, & 0 \le \tau < t,\ 0 \le |x| < r + s(t-\tau) - \varepsilon, \\ \frac{1}{\varepsilon}[r + s(t-\tau) - |x|], & 0 \le \tau < t,\ r + s(t-\tau) - \varepsilon \le |x| < r + s(t-\tau), \\ 0, & 0 \le \tau < t,\ r + s(t-\tau) \le |x| < \infty, \\ 0, & t \le \tau < \infty,\ x \in \mathbb{R}^m. \end{cases} \tag{A.4}$$

Then, by virtue of (1.4) and (1.7),

$$
\begin{aligned}
0 \geq & \int_0^\infty \int_{\mathbb{R}^m} \left\{ -\partial_t \phi \eta - \sum_{\alpha=1}^m \partial_\alpha \phi q^\alpha \right\} dx d\tau - \int_{\mathbb{R}^m} \phi(x,0)\eta(U_0(x))dx \\
= & \frac{1}{\varepsilon} \int_{t-\varepsilon}^t \int_{|x|<r} \eta(U(x,\tau))dx d\tau - \int_{|x|<r+st} \eta(U_0(x))dx \\
& + \frac{1}{\varepsilon} \int_0^t \int_{r+s(t-\tau)-\varepsilon<|x|<r+s(t-\tau)} \left\{ s\eta + \sum_{\alpha=1}^m \frac{x_\alpha}{|x|} q^\alpha \right\} dx d\tau + O(\varepsilon) \\
\geq & \frac{1}{\varepsilon} \int_{t-\varepsilon}^t \int_{|x|<r} \eta(U(x,\tau))dx d\tau - \int_{|x|<r+st} \eta(U_0(x))dx + O(\varepsilon).
\end{aligned} \tag{A.5}
$$

Since η is convex,

$$
\eta(U(x,\tau)) \geq \eta(U(x,t)) + D\eta(U(x,t))[U(x,\tau) - U(x,t)]. \tag{A.6}
$$

Therefore (A.5) yields

$$
\begin{aligned}
& \int_{|x|<r} \eta(U(x,t))dx - \int_{|x|<r+st} \eta(U_0(x))dx \\
& \leq \int_{|x|<r} D\eta(U(x,t))U(x,t)dx - \frac{1}{\varepsilon} \int_{t-\varepsilon}^t \int_{|x|<r} D\eta(U(x,t))U(x,\tau)dx d\tau + O(\varepsilon).
\end{aligned} \tag{A.7}
$$

Letting $\varepsilon \downarrow 0$ and using (A.1) we arrive at (1.8).

Note that any entropy–entropy flux pair $(\hat{\eta}, \hat{q})$ with $\hat{\eta}$ uniformly convex induces, through

$$
\begin{cases}
\eta(U) := \hat{\eta}(U) - \hat{\eta}(0) - D\hat{\eta}(0)U, \\
q^\alpha(U) := \hat{q}^\alpha(U) - \hat{q}^\alpha(0) - D\hat{\eta}(0)[F^\alpha(U) - F^\alpha(0)], \quad \alpha = 1, \ldots, m,
\end{cases} \tag{A.8}
$$

another entropy–entropy flux pair (η, q) which satisfies (1.7). Therefore L^2 stability (1.9) holds in that case.

REFERENCES

[1] P. Brenner, *The Cauchy problem for symmetric hyperbolic systems in L_p*, Math. Scand., 19 (1966), pp. 27–37.

[2] K. N. Chueh, C. C. Conley, and J. A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 372–411.

[3] J. Conlon and T.-P. Liu, *Admissibility criteria for hyperbolic conservation laws*, Indiana Univ. Math. J., 30 (1981), pp. 641–652.

[4] C. M. Dafermos, *Hyperbolic systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., D. Reidel, Dordrecht, the Netherlands, 1983, pp. 25–70.

[5] ———, *Quasilinear hyperbolic systems with involutions*, Arch. Rational Mech. Anal., 94 (1986), pp. 373–389.

[6] K. O. Friedrichs and P. D. Lax, *Systems of conservation equations with a convex extension*, Proc. Nat. Acad. Sci. U.S.A., 68 (1971), pp. 1686–1688.

[7] S. N. KRUZKOV, *First order quasilinear equations in several independent variables*, Mat. Sb. (N.S.), 81 (1970), pp. 228–255; Math. USSR-Sb., 10 (1970), pp. 217–243 (English Translation.)

[8] P. D. LAX, *Shock waves and entropy*, in Contributions to Functional Analysis, E.A. Zarantonello, ed., Academic Press, New York, 1971, pp. 603–634.

[9] J. RAUCH, *BV estimates fail for most quasilinear hyperbolic systems in dimensions greater than one*, Commun. Math. Phys., 106 (1986), pp. 481–484.

[10] D. SERRE, *Richness and the classification of quasilinear hyperbolic systems*, in Multidimensional Hyperbolic Problems and Computations, J. Glimm and A. Majda, eds., Springer-Verlag, New York, 1991, pp. 315-333.

[11] ———, *Domaines invariants pour les systèmes hyperboliques de lois de conservation*, J. Differential Equations, 69 (1987), pp. 46–62.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1415–1424, November 1995

002

A COMPARISON OF TWO VISCOUS REGULARIZATIONS OF THE RIEMANN PROBLEM FOR BURGERS'S EQUATION*

M. SLEMROD†

Abstract. This note compares solutions of two regularizations for Burgers' equation $u_t + (u^2/2)_x = 0$ with Riemann initial data $u = u_-(x \le 0)$, $u = u_+(x > 0)$ at $t = 0$. The regularizations are given by $u_t^\varepsilon + (u^{\varepsilon^2}/2)_x = \varepsilon u_{xx}^\varepsilon$ and $u_t^\varepsilon + (u^{\varepsilon^2}/2)_x = \varepsilon t u_{xx}^\varepsilon$ with appropriate initial data in each case. The first regularization is more traditional while the second preserves the space–time dilational invariance of the Riemann problem for the inviscid equation. Here it is shown that the difference of the two regularizations approaches zero (in appropriate integral norms depending on the data) as $\varepsilon \to 0_+$ for $0 < t \le 1$.

Key words. hyperbolic conservation law, viscous regularization

AMS subject classification. 35165

Introduction. The objective of this paper is to begin a comparison of two viscous regularizations of hyperbolic systems of conservation laws

$$\begin{aligned} &u_t + f(u)_x = 0, \quad \text{where } u : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^n , \\ &f'(u) \text{ possesses real distinct eigenvalues for all } u \in \mathbb{R}^n \end{aligned} \tag{0.1}$$

with Riemann initial data

$$u(x,0) = \begin{cases} u_- , & x \le 0 , \\ u_+ , & x > 0 . \end{cases} \tag{0.2}$$

The first regularization is quite classical: one simply imbeds (0.1) in the viscous equation

$$u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon u_{xx}^\varepsilon . \tag{0.3}$$

The difficulty with this regularization is that (0.3) does not possess space–time dilational invariance ($(x,t) \to (\alpha x, \alpha t)$, $\alpha > 0$) of (0.1), (0.2). That is, while solutions of the Riemann initial value problem (0.1), (0.2) depend only on the self-similar independent variable ξ, solutions of (0.3) will not. To circumvent this difficulty Dafermos [1], Kalasnikov [7], and Tupciev [11] independently suggested the viscous regularization of (0.1), (0.2) given by

$$v_t^\varepsilon + f(v^\varepsilon)_x = \varepsilon t v_{xx}^\varepsilon , \tag{0.4}$$

$$v^\epsilon(x,0) = \begin{cases} u_- , & x \le 0 , \\ u_+ , & x > 0 . \end{cases} \tag{0.5}$$

---

*Received by the editors August 12, 1993; accepted for publication (in revised form) March 3, 1994. This research was supported by National Science Foundation grants DMS-90006945 and INT-8914473 and Office of Naval Research grant N00014-93-1-0015.

†Center for Mathematical Sciences, University of Wisconsin-Madison, Madison, Wisconsin 53715.

Equations (0.4), (0:5) do possess the desired space–time dilational invariance. In fact substitution of the ansatz $v^\varepsilon(x,t) = \phi^\varepsilon(\xi)$ into (0.4), (0.5) yields the system of ordinary differential equations

$$-\xi\phi^{\varepsilon\prime} + f(\phi^\varepsilon)' = \varepsilon\phi^{\varepsilon\prime\prime}, \qquad -\infty < \xi < \infty \tag{0.6}$$

with boundary conditions

$$\phi^\varepsilon(-\infty) = u_-, \qquad \phi^\varepsilon(+\infty) = u_+, \tag{0.7}$$

and $' = \frac{d}{d\xi}$.

In his papers [1], [2] Dafermos gave a class of equations in the case $n = 2$ for which (0.6), (0.7) do indeed possess a solution and which converges, bounded almost everywhere, $-\infty < \xi < \infty$, to a solution of (0.1), (0.2). Shocks in his solution of (0.1), (0.2) are admissible according to the viscosity criterion, i.e., they are the limit of traveling wave solutions of (0.3). Further examples applying this method were given in the papers of Dafermos and DiPerna [3], Fan [4], Slemrod [8], and Slemrod and Tzavaras [9].

As mentioned above this note will compare the two regularizations of (0.1) given by (0.3) and (0.4). For simplicity only the canonical scalar case of Burgers's equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0 \tag{0.8}$$

with Riemann data (0.2) is considered here. The reason for this simplicity is well known: the solution of the Riemann problem for (0.8) possesses either a shock or a rarefaction wave, but not both.

For the scalar equation (0.4) with $n = 1$, it is an easy matter to see that (0.6), (0.7) possess uniformly bounded (in ξ, ε) solutions with uniformly bounded total variation. Hence $\{\phi^\varepsilon\}$ will contain a subsequence which converges bounded almost everywhere (a.e.) in ξ as $\varepsilon \to 0_+$ to a solution of Riemann problem (0.1), (0.2). Then the issue is as follows: Can we show that the error $u^\varepsilon - \phi^\varepsilon$ approaches zero in some sense for a class of data $u^\varepsilon(x,0)$ approaching Riemann data as $\varepsilon \to 0_+$? An affirmative answer is given for the case $f(u) = u^2/2$ in Theorems 1 and 2. This shows that the εt regularization may provide a convenient interpolation between inviscid equations (0.1) and fully viscous equations (0.3) for the case of Riemann problems. Of course a better test will come at the level of systems when $n = 2$, which still remains to be done.

The paper is divided into two sections after this Introduction. Section 1 considers the case $u_- \le u_+$ when the admissible solution of (0.1), (0.2) will be a continuous rarefaction wave. Section 2 considers the case $u_- > u_+$ when the admissible solution of (0.1), (0.2) will be a piecewise constant shock wave. Here admissible is taken to mean a solution satisfying the Lax entropy criterion (see, for example, [10]). The method of proof for both cases is modeled on an energy stability argument of Goodman [6] for stability of traveling wave solutions of (0.3). The difference here is that the viscosity is vanishing, while in [6] it remains fixed.

1. Rarefaction case $u_- \le u_+$. We consider (0.6), (0.7) for the case of Burgers's equation $f(u) = \frac{u^2}{2}$ and data $u_- \le u_+$. In this case (0.6), (0.7) becomes

$$-\xi\phi' + \left(\frac{\phi^2}{2}\right)' = \varepsilon\phi'', \tag{1.1}$$

$$\phi(-\infty) = u_-, \ \phi(+\infty) = u_+, \tag{1.2}$$

where the superscript ε has been suppressed. The theory of [1] can be applied to (1.1), (1.2) to obtain the existence of a solution. The standard uniqueness theorem for ordinary differential equations as applied to ϕ' shows that ϕ is monotone increasing when $u_- < u_+$ and ϕ is $u_-(= u_+)$ when $u_- = u_+$. Moreover, since $\phi' \geq 0$ and $\phi' \to 0$ as $|\xi| \to \infty$ we know ϕ' has a nonnegative maximum on $(-\infty, \infty)$. At such a maximum $\phi'' = 0$, $\phi''' \leq 0$, which from (1.1) implies $0 \leq \phi' \leq 1$. Hence solutions of (1.1), (1.2) are

$$\begin{aligned} &\text{(i) monotone increasing, } \ 0 < \phi' \leq 1 \ \text{ when } \ u_- < u_+ \ , \\ &\text{(ii) constant} = u_-(= u_+) \ \text{ when } \ u_- = u_+ \ . \end{aligned} \tag{1.3}$$

We will compare $v^\varepsilon(x,t) = \phi^\varepsilon(\frac{x}{t})$ with solutions $u^\varepsilon(x,t)$ of the viscous equation

$$u_t + \left(\frac{u^2}{2}\right)_x = \varepsilon u_{xx} \ , \quad t \geq t_0 > 0 \ , \quad -\infty < x < \infty, \tag{1.4}$$

$$u(x, t_0) = \phi\left(\frac{x}{t_0}\right) . \tag{1.5}$$

LEMMA 1. $\lim_{x \to +\infty} u = u_+$, $\lim_{x \to -\infty} u = u_-$ *for* $t > t_0$.

Proof. Differentiating (1.4) with respect to x and using the standard L^2 theory, one can prove the existence-uniqueness of smooth solutions of (1.4), (1.5), where $u_x \to 0$ as $|x| \to \pm\infty$. Next use the representation formula for the solution of the nonhomogeneous heat equation $u_t - \varepsilon u_{xx} = -(u^2/2)_x$ with data (1.5) as given by the representation formula in terms of the Green function for the heat operator (see, for example, [5]):

$$\begin{aligned} u(x,t) = (4\pi\varepsilon(t-t_0))^{-1/2} &\int_{-\infty}^{\infty} \phi\left(\frac{\xi}{t_0}\right) \exp(-(x-\xi)^2/4\varepsilon(t-t_0))d\xi \\ &- \frac{1}{2}\int_{-\infty}^{\infty}\int_{t_0}^{\infty} u_x^2(\xi,\tau)(4\pi\varepsilon(t-t_0))^{-1/2}\exp(-(x-\xi)^2/4\varepsilon(t-t_0-\tau))d\tau d\xi \ . \end{aligned}$$

The representation formula is the sum of two integral terms, one arising from the nonhomogeneous forcing $-(u^2/2)_x$ and the other from the initial data (1.5). We then let $|x| \to \infty$ in the two integrals and apply Laplace's method for the asymptotic evaluation of integrals. The fact that $u_x \to 0$ as $|x| \to \infty$ yields the second term zero as $|x| \to \infty$, while $\phi(\xi) \to u_+$ as $\xi \to +\infty$, $\phi(\xi) \to u_-$ as $\xi \to -\infty$ shows that the first term approaches u_+ as $x \to \infty$, u_- as $x \to -\infty$.

Define $y^\varepsilon(x,t) \doteq u^\varepsilon(x,t) - v^\varepsilon(x,t)$ so that y^ε satisfies

$$y_t + (f(u) - f(v))_x = \varepsilon y_{xx} + \varepsilon(1-t)v_{xx} \ , \tag{1.6}$$

$$y = 0 \ \text{ at } \ t = t_0 \ . \tag{1.7}$$

Next set

$$Y^\varepsilon(x,t) \doteq \int_{-\infty}^{x} y^\varepsilon(z,t)dz \ . \tag{1.8}$$

Since $u = u_+ = \phi(+\infty)$ at $x = \infty$, $u = u_- = \phi(-\infty)$ at $x = -\infty$ for $t > t_0$, we have

$$\frac{d}{dt}\int_{-\infty}^{\infty} y^\varepsilon dx - (f(u) - f(v))\Big|_{x=-\infty}^{x=+\infty} = 0\,,$$

and hence

$$\int_{-\infty}^{\infty} y^\varepsilon(x,t)dx = \int_{-\infty}^{\infty} y^\varepsilon(x,t_0)dx = 0 \quad \text{for} \quad t \geq t_0\,, \tag{1.9}$$

so that

$$Y^\varepsilon \to 0 \quad \text{as} \quad |x| \to \infty, \quad t \geq t_0\,, \tag{1.10}$$

while

$$Y^\varepsilon = 0 \quad \text{at} \quad t = t_0\,, \quad -\infty < x < \infty\,. \tag{1.11}$$

Also, substitution of $Y_x^\varepsilon = y^\varepsilon$ into (1.6) yields

$$Y_{xt} + \frac{1}{2}((Y_x + u)^2 - (v)^2)_x = \varepsilon Y_{xxx} + \varepsilon(1-t)v_{xx},$$

which upon integration from $-\infty$ to x and application of (1.11) shows that Y satisfies

$$Y_t + vY_x + \frac{Y_x^2}{2} = \varepsilon Y_{xx} + \varepsilon(1-t)v_x, \qquad t > t_0\,, \tag{1.12}$$

and boundary and initial conditions (1.10), (1.11).

We can now state and prove a theorem comparing the two viscous regularizations in the case when $u_- \leq u_+$.

LEMMA 2. *For the case* $f(u) = u^2/2$, $u_- \leq u_+$, *set* $Y^\varepsilon(x,t;t_0) = \int_{-\infty}^{x}(u^\varepsilon(z,t) - v^\varepsilon(z,t))dz$, *where* u^ε *satisfies* (1.4), (1.5) *and* $v^\varepsilon(x,t) = \phi^\varepsilon(\frac{x}{t}) = \phi^\varepsilon(\xi)$ *satisfies* (1.1), (1.2). *Then*

$$0 \leq Y^\varepsilon(x,t;t_0) \leq \varepsilon\Big(\ln\Big(\frac{t}{t_0}\Big) - (t - t_0)\Big) \quad \text{on} \quad t_0 \leq t \leq 1\,.$$

Proof. Since $v^\varepsilon(x,t) = \phi^\varepsilon(\frac{x}{t})$ and $v_x^\varepsilon(x,t) = \phi^{\varepsilon\prime}(\frac{x}{t})\frac{1}{t}$ with $0 \leq \phi^{\varepsilon\prime} \leq 1$, we know $0 \leq v_x^\varepsilon(x,t) \leq \frac{1}{t}$ and $0 \leq \varepsilon(1-t)v_x^\varepsilon \leq \varepsilon(\frac{1}{t} - 1)$ for $t_0 \leq t \leq 1$. It then follows that $\underline{Y} = 0$ is a subsolution and $\bar{Y} = \varepsilon(\ln\ t - t) - \varepsilon(\ln\ t_0 - t_0)$ is a supersolution of (1.11)–(1.12) on $t_0 \leq t \leq 1$. Hence

$$0 \leq Y^\varepsilon(x,t;t_0) \leq \varepsilon\Big(\ln\Big(\frac{t}{t_0}\Big) - (t - t_0)\Big), \tag{1.13}$$

and the lemma follows immediately.

THEOREM 1. *For the case* $f(u) = u^2/2$, $u_- \leq u_+$,

(i) $$\int_{t_0}^{\bar{t}} \frac{1}{t}\int_{-\infty}^{\infty}(u^\varepsilon(x,t) - v^\varepsilon(x,t))^2 dxdt \leq \frac{\varepsilon}{2}\Big(\ln\Big(\frac{\bar{t}}{t_0}\Big) - (\bar{t} - t_0)\Big)^2(u_+ - u_-) \quad \text{on} \quad t_0 \leq \bar{t} \leq 1;$$

in particular with $t_0 = \varepsilon$, $\bar{t} = 1$,

$$\text{(ii)} \qquad \lim_{\varepsilon \to 0+} \int_{\varepsilon}^{1} \frac{1}{t} \int_{-\infty}^{\infty} (u^{\varepsilon}(x,t) - v^{\varepsilon}(x,t))^2 dx dt = 0.$$

Proof. From (1.13) we know $Y^{\varepsilon} \geq 0$, hence multiplication of (1.12) by $Y = Y^{\varepsilon}$ and integration by parts yields

$$\frac{1}{2}\frac{d}{dt}\int_{-\infty}^{\infty} Y^2 dx - \frac{1}{2}\int_{-\infty}^{\infty} v_x Y^2 dx + \varepsilon \int_{-\infty}^{\infty} Y_x^2 dx \leq \varepsilon(1-t)\int_{-\infty}^{\infty} Y v_x dx \,. \tag{1.14}$$

Since $0 \leq \phi' \leq 1$ we know $0 \leq v_x \leq \frac{1}{t}$ and (1.14) implies

$$\frac{1}{2}\frac{d}{dt}\int_{-\infty}^{\infty} Y^2 dx - \frac{1}{2t}\int_{-\infty}^{\infty} Y^2 dx + \varepsilon \int_{-\infty}^{\infty} Y_x^2 dx \leq \varepsilon(1-t)\int_{-\infty}^{\infty} Y v_x dx$$

or

$$\frac{1}{2}\frac{d}{dt}\Big(\frac{1}{t}\int_{-\infty}^{\infty} Y^2 dx\Big) + \frac{\varepsilon}{t}\int_{-\infty}^{\infty} Y_x^2 dx \leq \varepsilon\Big(\frac{1}{t} - 1\Big)\int_{-\infty}^{\infty} Y v_x dx \tag{1.15}$$

for $t_0 \leq t \leq 1$. But on this interval we know from Lemma 2 that

$$0 \leq Y^{\varepsilon}(x,t) \leq \varepsilon L(t), \qquad L(t) \doteq \ln\Big(\frac{t}{t_0}\Big) - (t - t_0). \tag{1.16}$$

Hence (1.16), when substituted in (1.15), gives

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\Big(\frac{1}{t}\int_{-\infty}^{\infty} Y^2 dx\Big) + \frac{\varepsilon}{t}\int_{-\infty}^{\infty} Y_x^2 dx &\leq \varepsilon L'(t) L(t) \int_{-\infty}^{\infty} v_x dx \\ &= \frac{\varepsilon}{2}\frac{d}{dt} L^2(t)(u_+ - u_-) \,. \end{aligned} \tag{1.17}$$

Now integrate (1.17) from t_0 to $\bar{t}$ using initial condition (1.11), and the theorem is proven.

Remark 1. It is result (ii) that gives the sharp estimate on how the two regularizations agree. This is true because the initial data (1.5), $u(x,\varepsilon) = \phi(\frac{x}{\varepsilon})$ approaches the Riemann data: $\phi(\frac{x}{\varepsilon}) \to u_-$ $x < 0$, $\phi(\frac{x}{\varepsilon}) \to u_+$, $x > 0$. Note that we needed to use data (1.5) and not the Riemann data to enforce the initial condition (1.11).

Remark 2. The restriction $0 < t \leq 1$ makes sense: for t large $\sim \frac{1}{\varepsilon}$, (0.3) will approach (0.1) as $\varepsilon \to 0_+$, while (0.4) will remain parabolic.

2. Shock case $u_- > u_+$. In the case $u_- > u_+$, ϕ the solution of (1.1), (1.2) is of course monotone decreasing with no uniform (in ε) bound from below on ϕ'. Hence a different method of analysis from §1 is required.

First we introduce the traveling wave variable σ and the rescaled time variable $\tau : \sigma = \frac{x - st}{\varepsilon}$, $\tau = \frac{t}{\varepsilon}$. In these variables (0.3) and (0.4) become, respectively,

$$u_\tau - s u_\sigma + \Big(\frac{u^2}{2}\Big)_\sigma = u_{\sigma\sigma} \tag{2.1}$$

and

$$v_\tau - s v_\sigma + \Big(\frac{v^2}{2}\Big)_\sigma = \varepsilon \tau v_{\sigma\sigma} \,. \tag{2.2}$$

Equation (2.1) admits the one parameter family of traveling wave solutions $u(\sigma,\tau) = \psi(\sigma-\sigma_0)$ independent of τ satisfying

$$-s(\psi-u_-)+\frac{\psi^2-u_-^2}{2}=\psi' , \tag{2.3}$$

$$\psi(-\infty)=u_-,\ \psi(+\infty)=u_+ , \tag{2.4}$$

$$s=\frac{1}{2}(u_++u_-) . \tag{2.5}$$

For $v^\varepsilon(x,t)$ we once again impose Riemann data (0.5) so that $v^\varepsilon(x,t)=\phi^\varepsilon(\frac{x}{t})$, where $\phi^\varepsilon(\xi)$ satisfies (0.6).

We choose σ_0 so that

$$\int_{-\infty}^{-\sigma_0}\{\psi(x)-u_-\}\,dx+\int_{-\sigma_0}^{+\infty}\{\psi(x)-u_+\}\,dx=0 . \tag{2.6}$$

This can be done because ψ is monotone decreasing with $\psi(-\infty)=u_-$, $\psi(+\infty)=u_+$, and the left-hand side of (2.6) approaches $-\infty(+\infty)$ as σ_0 approaches $-\infty(+\infty)$. Hence there is some intermediate value σ_0 for which (2.6) holds.

As in §1 we define

$$y^\varepsilon(\sigma,\tau)=u(\sigma,\tau)-v^\varepsilon(\sigma,\tau)=\psi(\sigma-\sigma_0)-v^\varepsilon(\sigma,\tau) \tag{2.7}$$

and let

$$Y^\varepsilon(\sigma,\tau)=\int_{-\infty}^{\sigma}\{\psi(\sigma-\sigma_0)-v^\varepsilon(\sigma,\tau)\}\,d\sigma . \tag{2.8}$$

We see y^ε satisfies

$$y_\tau-sy_\sigma+\Big(\frac{u^2}{2}-\frac{v^2}{2}\Big)_\sigma=\varepsilon\tau y_{\sigma\sigma}+(1-\varepsilon\tau)u_{\sigma\sigma} . \tag{2.9}$$

Integration of (2.9) shows

$$\frac{d}{d\tau}\int_{-\infty}^{\infty}y(\sigma,\tau)d\sigma-sy\Big|_{\sigma=-\infty}^{\sigma=+\infty}+\Big(\frac{u^2}{2}-\frac{v^2}{2}\Big)\Big|_{\sigma=-\infty}^{\sigma=+\infty}=0 .$$

But for $0<\tau$ fixed $\sigma=\pm\infty$ implies $x=\pm\infty$ and $u(\infty,\tau)=\psi(+\infty)=\phi(+\infty)=u_+$, $u(-\infty,\tau)=\psi(-\infty)=\psi(-\infty)=u_-$, and hence

$$\int_{-\infty}^{\infty}y(\sigma,\tau)d\sigma=\int_{-\infty}^{\infty}y(\sigma,0)d\sigma .$$

Now write

$$\begin{aligned}\int_{-\infty}^{\infty}y(\sigma,0)d\sigma&=\int_{-\infty}^{\infty}\{\psi(\sigma-\sigma_0)-v^\varepsilon(\sigma,0)\}\,d\sigma\\&=\int_{-\infty}^{0}\{\psi(\sigma-\sigma_0)-u_-\}\,d\sigma+\int_{0}^{\infty}\{\psi(\sigma-\sigma_0)-u_+\}\,d\sigma\\&=\int_{-\infty}^{-\sigma_0}\{\psi(x)-u_-\}\,dx+\int_{-\sigma_0}^{\infty}\{\psi(x)-u_+\}\,dx=0\end{aligned}$$

by (2.6) so that

$$\int_{-\infty}^{\infty} y(\sigma,\tau)d\sigma = 0 \quad \text{for} \quad \tau \geq 0,$$

and hence

$$Y^\varepsilon(\sigma,\tau) \to 0 \quad \text{as} \quad |\sigma| \to \infty, \qquad \tau \geq 0\,. \tag{2.10}$$

Next substitute $Y^\varepsilon_\sigma = y^\varepsilon$ into (2.9) to obtain

$$Y_{\sigma\tau} - sY_{\sigma\sigma} + \left(\frac{u^2}{2} - \frac{v^2}{2}\right)_\sigma = \varepsilon\tau Y_{\sigma\sigma\sigma} + (1-\varepsilon\tau)u_{\sigma\sigma},$$

which upon integration from $-\infty$ to σ yields

$$Y_\tau - sY_\sigma + uY_\sigma - \frac{Y_\sigma^2}{2} = \varepsilon\tau Y_{\sigma\sigma} + (1-\varepsilon\tau)u_\sigma\,. \tag{2.11}$$

At $\tau = 0$ we see from (2.6) that

$$Y^\varepsilon(\sigma,0) = \int_{-\infty}^{\sigma} \{\psi(\sigma-\sigma_0) - v^\varepsilon(\sigma,0)\}\, d\sigma = \frac{1}{\varepsilon}\int_{-\infty}^{x} \left\{\psi\left(\frac{z}{\varepsilon} - \sigma_0\right) - v_-\right\} dz \leq 0$$

if $\sigma = \frac{x}{\varepsilon} \leq 0$;

$$Y^\varepsilon(\sigma,0) = \frac{1}{\varepsilon}\int_{x}^{\infty} \left\{\psi\left(\frac{z}{\varepsilon} - \sigma_0\right) - v_+\right\} dz \leq 0 \quad \text{if} \quad \sigma = \frac{x}{\varepsilon} > 0.$$

Hence we know $Y^\varepsilon(\sigma,\tau)$ satisfies (2.11), (2.10) and

$$Y^\varepsilon(\sigma,0) \leq 0, \qquad -\infty < \sigma < \infty\,. \tag{2.12}$$

We are now in a position to state and prove a theorem comparing the two viscous regularizations when $u_- > u_+$.

THEOREM 2. *For the case* $f(u) = u^2/2$, $u_- > u_+$ *let* $u^\varepsilon(x,t) = \psi(\sigma - \sigma_0)$, *where* ψ *is the traveling wave solution of* (0.3), *i.e.,* ψ, σ_0 *satisfy* (2.3)–(2.6). *Let* $v^\varepsilon(x,t) = \phi^\varepsilon(\frac{x}{t}) = \phi^\varepsilon(\xi)$ *be the solution of* (0.4), (0.5) *where* ϕ^ε *satisfies* (1.1), (1.2). *Then there exists a const. (independent of* ε*) depending only on* u_-, u_+, *so that for* $0 < \bar{t} \leq 1$,

$$\int_0^{\bar{t}}\int_{-\infty}^{\infty} t(u^\varepsilon(x,t) - v^\varepsilon(x,t))^2 dxdt \leq \text{const.}\ \varepsilon\,;$$

in particular, with $\bar{t} = 1$,

$$\lim_{\varepsilon\to 0}\int_0^1\int_{-\infty}^{\infty} t(u^\varepsilon(x,t) - v^\varepsilon(x,t))^2 dxdt = 0\,.$$

Remark 3. Note that the initial data for u^ε given by $u^\varepsilon(x,0) = \psi(\frac{x}{\varepsilon} - \sigma_0)$ approaches the Riemann data (0.2) as $\varepsilon \to 0_+$.

Remark 4. As noted in Remark 2 the restriction $0 \leq \bar{t} \leq 1$ is natural.

Proof. Since $u_\sigma = \psi'(\sigma - \sigma_0) < 0$, $\bar{Y} = 0$ is a supersolution of (2.10)–(2.12) when $0 \le 1 - \varepsilon\tau$, $0 < \tau$, i.e., $0 < t \le 1$. Hence $Y^\varepsilon \le 0$ or $0 < \tau < \frac{1}{\varepsilon}$, $-\infty < \sigma < \infty$. Now multiply (2.11) by Y^ε. We obtain

$$\frac{1}{2}\frac{d}{d\tau}\int_{-\infty}^{\infty} Y^2\,d\sigma + s\int_{-\infty}^{\infty}\Big(\frac{Y^2}{2}\Big)_\sigma d\sigma + \int_{-\infty}^{\infty} u\Big(\frac{Y^2}{2}\Big)_\sigma - \int_{-\infty}^{\infty}\frac{Y_\sigma^2 Y}{2}d\sigma$$
$$= -\varepsilon\tau\int_{-\infty}^{\infty} Y_\sigma^2 d\sigma + (1-\varepsilon\tau)\int_{-\infty}^{\infty} u_\sigma Y\,d\sigma,$$

which upon integration by parts becomes

$$\frac{1}{2}\frac{d}{d\tau}\int_{-\infty}^{\infty} Y^2 d\sigma - \int_{-\infty}^{\infty} u_\sigma\Big(\frac{Y^2}{2}\Big)d\sigma - \int_{-\infty}^{\infty}\frac{Y_\sigma^2 Y}{2}d\sigma$$
$$= -\varepsilon\tau\int_{-\infty}^{\infty} Y_\sigma^2 d\sigma + (1-\varepsilon\tau)\int_{-\infty}^{\infty} u_\sigma Y\,d\sigma. \tag{2.13}$$

But since $Y \le 0$, the third term on the left-hand side of (2.13) is nonnegative, and hence we have

$$\frac{1}{2}\frac{d}{d\tau}\int_{-\infty}^{\infty} Y^2 d\sigma + \varepsilon\tau\int_{-\infty}^{\infty} y^2 d\sigma$$
$$\le \int_{-\infty}^{\infty}\psi'(\sigma-\sigma_0)\Big(\frac{Y^2}{2} + (1-\varepsilon\tau)Y\Big)d\sigma \tag{2.14}$$
$$\le \int_{-\infty}^{\infty}\frac{\psi'(\sigma-\sigma_0)}{2}\{(Y+(1-\varepsilon\tau))^2 - (1-\varepsilon\tau)^2\}d\sigma\,.$$

But $\psi' < 0$, and hence (2.14) implies

$$\frac{1}{2}\frac{d}{d\tau}\int_{-\infty}^{\infty} Y^2 d\sigma + \varepsilon\tau\int_{-\infty}^{\infty} y^2 d\sigma \le \frac{(1-\varepsilon\tau)^2}{2}(u_- - u_+) \le \frac{(u_- - u_+)}{2}\,. \tag{2.15}$$

Before proceeding further we shall estimate the L^2 norm of $Y(\sigma, 0)$.

$$Y^\varepsilon(\sigma,0) = \int_{-\infty}^{\sigma}\{\psi(\sigma-\sigma_0) - v^\varepsilon(\sigma,0)\}\,d\sigma \quad \text{so that}$$
$$Y^\varepsilon(\sigma,0) = \int_{-\infty}^{\sigma}\{\psi(\sigma-\sigma_0) - u_-\}\,d\sigma \quad \text{for} \quad \sigma \le 0$$
$$= -\int_{\sigma}^{\infty}\{\psi(\sigma-\sigma_0) - u_+\}\,d\sigma \quad \text{for} \quad \sigma > 0\,.$$

Of course we could substitute the explicit formula $\psi(\sigma) = u_+ + \frac{1}{2}(u_+ - u_-)[1 - \tanh((u_- - u_+)\sigma/4)]$ at this stage and obtain the desired estimate. However, with possible generalizations in mind for more general flux functions and systems, it seems better to obtain the bound without special formulas.

For notational convenience set $Y^\varepsilon(\sigma,0) = Y(\sigma)$ for the moment. Then, on $-\infty < \sigma \le 0$, (2.3) implies

$$Y''(\sigma) + h(\sigma)Y'(\sigma) = 0 \tag{2.16}$$

with $h(\sigma) = s-((\psi + u_-)/2) = (u_+-\psi)/2 \geq k > 0$. Equation (2.16) implies that $Y(\sigma)$ satisfies $Y(\sigma) = Y_\sigma(0)\int_{-\infty}^{\sigma} \exp(-\int_0^{\bar\sigma} hd)d\bar\sigma \leq \psi_\sigma(0)\int_{-\infty}^{\sigma} e^{k\bar\sigma}d\bar\sigma \leq (e^{k\sigma}/k)\psi_\sigma(0-) \leq (e^{k\sigma}/k)(\psi(-\sigma_0) - u_-)$ on $(-\infty, 0]$. A similar bound holds on $[0, \infty)$, and hence

$$\int_{-\infty}^{\infty} Y^\varepsilon(\sigma, 0)^2 d\sigma \leq \text{const.},$$

where const. is independent of ε.

Next integrate (2.15) from $\tau = 0$ to $\tau(\leq \frac{1}{\varepsilon})$ to obtain

$$\begin{aligned} &\frac{1}{2}\int_{-\infty}^{\infty} Y^2(\sigma,\tau)d\sigma + \varepsilon\int_0^{\tau}\int_{-\infty}^{\infty} \bar\tau y^2(\sigma,\bar\tau)d\sigma d\bar\tau \\ &\quad\leq \frac{\tau}{2}(u_- - u_+) + \frac{1}{2}\int_{-\infty}^{\infty} Y^2(\sigma,0)d\sigma \\ &\quad\leq \frac{\tau}{2}(u_- - u_+) + \text{const.}, \end{aligned} \tag{2.17}$$

where the dependence of y, Y on ε is again suppressed.

We now switch back to the independent variables x, t and denote

$$\tilde y(x,t) = y(\sigma,\tau) = u^\varepsilon(x,t) - v^\varepsilon(x,t), \text{where } u^\varepsilon(x,t) = \psi\Big(\frac{x-st}{\varepsilon} - \sigma_0\Big)$$

and $v^\varepsilon(x,t)$ is the solution of (0.4), (0.5). We see from (2.17) that for $0 < \bar t \leq 1$,

$$\varepsilon\int_0^{\bar t}\int_{-\infty}^{\infty} \frac{t\,\tilde y^2(x,t)\,dx\,dt}{\varepsilon \quad\quad \varepsilon\ \varepsilon} \leq \frac{\bar t}{2\varepsilon}(u_- - u_+) + \text{const},$$

i.e.,

$$\begin{aligned} \int_0^{\bar t}\int_{-\infty}^{\infty} t(u^\varepsilon(x,t) - v^\varepsilon(x,t))^2 dxdt &\leq \varepsilon\bar t(u_- - u_+) + \text{const. } \varepsilon \\ &\leq \text{const. } \varepsilon. \end{aligned} \tag{2.18}$$

Inequality (2.18) yields the desired result.

Acknowledgment. The author thanks Professors A. Tzavaras and M. Rascle and the referee for their valuable remarks regarding this work.

REFERENCES

[1] C. M. DAFERMOS, *Solution of the Riemann problem for a class of hyperbolic conservation laws by the viscosity method*, Arch. Rational Mech. Anal., 52 (1973), pp. 1–9.

[2] ———, *Structure of solutions of the Riemann problem for hyperbolic systems of conservation laws*, Arch. Rational Mech. Anal., 53 (1974), pp. 203–217.

[3] C. M. DAFERMOS AND R. J. DIPERNA, *The Riemann problem for certain classes of hyperbolic systems of conservations laws*, J. Differential Equations, 20 (1976), pp. 90–114.

[4] H. FAN, *A limiting "viscosity" approach to the Riemann problem for materials exhibiting a change of phase* (II), Arch. Rational Mech. Anal., 116 (1991), pp. 317–337.

[5] G. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[6] J. GOODMAN, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws*, Arch. Rational Mech. Anal., 95 (1986), pp. 325–344.

[7] A. S. KALASNIKOV, *Construction of generalized solutions of quasi-linear equations of first order without convexity conditions as limits of parabolic equations with a small parameter*, Dokl. Akad. Nauk. SSSR, 127 (1959), pp. 27–30. (In Russian.)

[8] M. SLEMROD, *A limiting "viscosity" approach to the Riemann problem for materials exhibiting change of phase*, Arch. Rational Mech. Anal., 105 (1989), pp. 327–365.

[9] M. SLEMROD AND A. E. TZAVARAS, *A limiting viscosity approach for the Riemann problem in isentropic gas dynamics*, Indiana Univ. Math. J., 38 (1989), pp. 1047–1074.

[10] J. SMOLLER, *Shock Waves and Reaction Diffusion Equations*, Springer-Verlag, New York, Berlin, Heidelberg, 1983.

[11] V. A. TUPCIEV, *On the method of introducing viscosity in the study of problems involving the decay of a discontinuity*, Dokl. Akad. Nauk. SSSR, 211 (1973), pp. 55–58; Soviet Math. Dokl., 14 (1973), pp. 978=–982. (English translation.)

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1425–1451, November 1995

003

ON SCALAR CONSERVATION LAWS WITH POINT SOURCE AND DISCONTINUOUS FLUX FUNCTION*

STEFAN DIEHL†

Abstract. The conservation law studied is $\frac{\partial u(x,t)}{\partial t} + \frac{\partial}{\partial x}(F(u(x,t),x)) = s(t)\delta(x)$, where u is a concentration, s is a source, δ is the Dirac measure, and

$$F(u,x) = \begin{cases} f(u), & x > 0, \\ g(u), & x < 0 \end{cases}$$

is the flux function. The special feature of this problem is the discontinuity that appears along the t-axis and the curves of discontinuity that go into and emanate from it. Necessary conditions for the existence of a piecewise smooth solution are given. Under some regularity assumptions sufficient conditions are given enabling construction of piecewise smooth solutions by the method of characteristics. The selection of a unique solution is made by a coupling condition at $x = 0$, which is a generalization of the classical entropy condition and is justified by studying a discretized version of the problem by Godunov's method.

The motivation for studying this problem is the fact that it arises in the modelling of continuous sedimentation of solid particles in a liquid.

Key words. conservation laws, discontinuous flux, point source

AMS subject classifications. 35A07, 35L65, 35Q80, 35R05

1. Preliminaries.

1.1. Introduction. This paper is a shortened version of [7], to which we refer for further details.

Let $u(x,t)$ be a scalar function, describing some kind of density, of the space coordinate x and the time coordinate t. It is well known that solutions of the initial value problem for a nonlinear scalar conservation law

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

where $f \in C^2$, even for $u_0 \in C^\infty$, may form discontinuities after a finite time. By interpreting the problem in a weak sense it is possible to define global discontinuous solutions. Uniqueness is guaranteed by an entropy condition. If f is nonconvex, the behaviour of the discontinuities is more complicated than in the convex case; see, e.g., Ballou [1]. He uses the method of characteristics to construct piecewise smooth solutions for piecewise constant initial data and "admissible" initial data (see the definition in [1]). Cheng [4] uses another method to construct solutions for bounded and piecewise monotone initial data. Dafermos [5] has shown that weak solutions of (1.1) generically are piecewise smooth if f has one inflection point.

Motivated by a part of the modelling of continuous sedimentation of solid particles in a liquid, to be described shortly in §1.3, we shall study a more general conservation problem with a point source and a discontinuous flux function. The problem will be described in §1.2. The questions of existence and uniqueness will be analysed in §2, which contains the main results: Theorem 2.17 on existence and Theorems 2.18, 2.19, and 2.20 on uniqueness. The solutions are selected by means of a coupling condition,

* Received by the editors January 11, 1993; accepted for publication March 4, 1994.
† Department of Mathematics, Lund Institute of Technology, P.O. Box 118, S-221 00 Lund, Sweden.

Condition Γ, which generalizes the classical entropy condition (Proposition 2.9). In §3 Condition Γ is numerically justified by studying a discretized conservation problem obtained by a scheme of Godunov type. The equivalence between Condition Γ and the so called *viscous profile condition* is analysed in [8]. The stability of these viscous profiles is studied in [9].

A special case of our problem is the Riemann problem with discontinuous flux function, dealt with in §2.3. In this problem there is no source term and the initial value is simple. The problem was addressed earlier by Gimse and Risebro [11]. In [12] Gimse and Risebro have, by construction of a sequence of approximate solutions, proved the existence of a solution of the Cauchy problem for a conservation law with discontinuous flux function arising in two-phase flow. They have left the question of uniqueness open. A uniqueness result for this type of equation is given at the end of §2.5. The presence of a point source causes considerable complications, even if there is no discontinuity in the flux function. Liu [17] studies nonlinear resonance when the source also depends on the state variable u. Another related problem is the initial boundary value problem in the sense of Bardos, Le Roux, and Nedelec [2]. In that problem a discontinuity is allowed along the boundary as long as it would like to propagate *out from* the domain. Confining our problem to one quadrant (of the x-t-plane), we get an initial value boundary flux problem; see §2.6. In this problem the flux at the boundary is prescribed and a value at the boundary is allowed to produce a discontinuity if and only if this discontinuity propagates *into* the domain.

1.2. The problem and assumptions. Let $s(t)$ be a source situated at $x = 0$, where the flux function $F(u,x)$ is a discontinuous function of x. Given initial data $u(x,0) = u_0(x)$, $x \in \mathbb{R}$, the weak formulation of the problem is

(1.2)
$$\int_0^\infty \int_{-\infty}^{\infty} (u\varphi_t + F\varphi_x)\,dx\,dt + \int_{-\infty}^{\infty} u_0(x)\varphi(x,0)\,dx + \int_0^\infty s(t)\varphi(0,t)\,dt = 0, \quad \varphi \in C_0^\infty(\mathbb{R}^2),$$

$$\text{where} \quad F(u,x) = \begin{cases} f(u), & x > 0, \\ g(u), & x < 0. \end{cases}$$

In the distribution sense it can be written

$$\begin{aligned} &\frac{\partial u(x,t)}{\partial t} + \frac{\partial}{\partial x}(F(u(x,t),x)) = s(t)\delta(x), && x \in \mathbb{R},\ t > 0, \\ &u(x,0) = u_0(x), && x \in \mathbb{R}, \end{aligned}$$

where $\delta(x)$ is the Dirac measure. If u is a smooth function except along $x = 0$, then by standard arguments it is easy to show that (1.2) is equivalent to

(1.3)
$$\begin{aligned} &u_t + f(u)_x = 0, && x > 0,\ t > 0, \\ &u_t + g(u)_x = 0, && x < 0,\ t > 0, \\ &f\big(u(0+,t)\big) = g\big(u(0-,t)\big) + s(t), && t > 0, \\ &u(x,0) = u_0(x), && x \in \mathbb{R}. \end{aligned}$$

The weak formulation (1.2) can allow a Dirac measure in u. For example, if $f \equiv$ constant $= f_0$ and $g \equiv$ constant $= g_0$, then $u(x,t) = u_0(x) + \delta(x)\int_0^t \big(s(\tau) + g_0 - f_0\big)\,d\tau$ is a solution. To avoid this, by a *solution* we mean a function u satisfying (1.2). A

function u is said to be *piecewise smooth* if it is bounded and C^1 except along a finite number of C^1-curves in every bounded set, such that the left and right limits of u along discontinuity curves exist. We especially introduce the notation

$$u_\pm(t) = \lim_{\delta \searrow 0} u(\pm\delta, t),$$
$$u^\pm(t) = \lim_{\varepsilon \searrow 0} u_\pm(t+\varepsilon).$$

The order of the limit processes is significant, for example, when a discontinuity reaches the t-axis or $s(t)$ is discontinuous. Note that $u^\pm(t)$ are continuous from the right. A function of one variable is said to be *piecewise monotone* if there are at most a finite number of points on every bounded interval where a shift of monotonicity occurs. We define a *discrete set* of real numbers as a set that contains at most a finite number of points on every bounded interval.

Assumptions. In this paper problem (1.2) will not be treated in full generality. To motivate a uniqueness condition for the discontinuity along the t-axis, the analysis is restricted to solutions in the class

$$\Sigma = \{u = u(x,t) : u \text{ is piecewise smooth, } u^\pm(t) \text{ are piecewise monotone}\}.$$

The initial value function u_0 is assumed to be piecewise monotone and piecewise smooth. The source function s is assumed to be piecewise monotone, piecewise smooth with bounded derivative, and continuous from the right. The flux functions $f, g \in C^2$ are assumed to have at most a discrete number of stationary points and the property $|f(u)|, |g(u)| \to \infty$ as $|u| \to \infty$. The last assumption is made to avoid unbounded solutions; see an example in [7]. To be able to construct a solution of problem (1.2), we assume that it is *regular* in a sense defined in §2.4.

1.3. Physical motivation. *Continuous sedimentation* of solid particles in a liquid takes place in a clarifier-thickener unit or settler; see Fig. 1.1. The one-dimensional

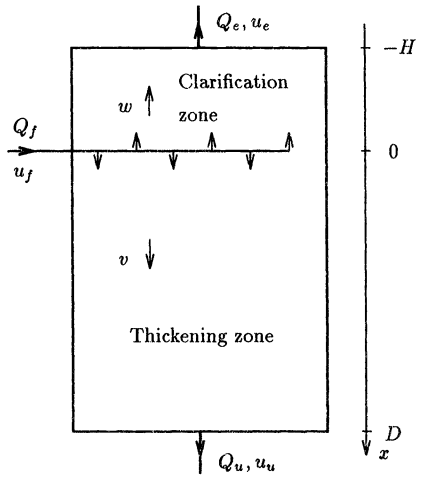


FIG. 1.1. *Schematic picture of the continuous clarifier-thickener.*

x-axis is shown in the figure. The height of the clarification zone is denoted by H and the depth of the thickening zone is denoted by D. At $x = 0$ the settler is fed

with suspended solids at a concentration $u_f(t)$ and at a constant flow rate Q_f (volume per unit time). A high concentration of solids is taken out at the underflow at $x = D$ at a rate Q_u. The effluent flow Q_e at $x = -H$ is consequently defined by $Q_e = Q_f - Q_u$. It is assumed that these three flows are positive. The cross-sectional area A is assumed to be constant and the concentration u is assumed to be constant on each cross section. We define the bulk velocities in the thickening and clarification zone as $v = Q_u/A$ and $w = Q_e/A$ with directions shown in Fig. 1.1. The feed inlet is modelled by the source function $s(t) = Q_f u_f(t)/A \geq 0$ (mass per unit area and unit time). The standard batch settling flux $\phi(u)$, introduced by Kynch [15] and still used today, is shown in Fig. 1.2, where $u_{\max}$ is the maximal packing concentration and u_{infl} is an inflection point. The phenomena at the feed level may be modelled by equations (1.3) with the flux functions $f(u) = \phi(u) + vu$ and $g(u) = \phi(u) - wu$. The theory of this paper can also be used to predict the effluent concentrations $u_e(t)$ and $u_u(t)$. An analysis of the sedimentation problem is carried out in [6].

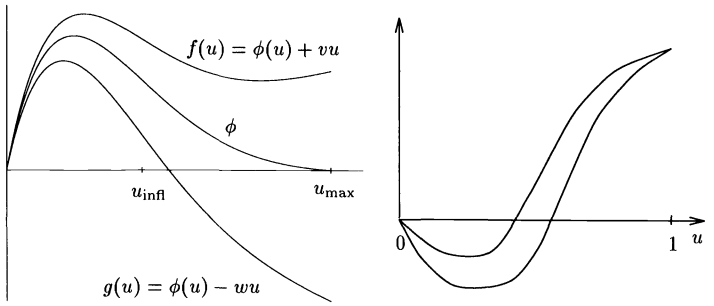


FIG. 1.2. *The flux curves in the sedimentation problem (left) and in the problem of two-phase flow (right). Note that in the right figure either of the graphs can be f or g.*

Another context where a discontinuous flux function appears is in the modelling of *two-phase flow through one-dimensional porous media*; see Gimse and Risebro [12] and the references therein. The source function is then $s \equiv 0$ and the flux function F may have several discontinuities in the space coordinate. The qualitative behaviour of these discontinuities may be analysed by letting the flux functions f and g in (1.3) have the shapes as shown in Fig. 1.2 with one global minimum, $f(0) = g(0)$ and $f(1) = g(1)$. At the end of §2.5 the questions of existence and uniqueness for the two problems when f is the upper (lower) and g is the lower (upper) curve in Fig. 1.2 (right) are commented upon.

1.4. Properties of nonconvex scalar conservation laws. In this section we review some basic properties of the solution of the scalar problem

$$\begin{aligned} &u_t + f(u)_x = 0, \quad x \in \mathbb{R},\ t > 0, \\ &u(x,0) = u_0(x), \quad x \in \mathbb{R}. \end{aligned} \tag{1.4}$$

If $x = x(t)$ is a C^1-curve of discontinuity for u, it obeys the *jump condition* or *Rankine–Hugoniot condition*

$$x'(t) = S(u^{x+}, u^{x-}),$$

where $S(\alpha,\beta) = \frac{f(\alpha)-f(\beta)}{\alpha-\beta}$ for $\alpha \neq \beta$ and $u^{x\pm} = u\big(x(t)\pm 0, t\big)$. Unstable discontinuities are rejected by imposing the *entropy condition*

$$S(v, u^{x-}) \geq S(u^{x+}, u^{x-}) \quad \forall v \text{ between } u^{x-} \text{ and } u^{x+} \tag{1.5}$$

by Oleinik [18]. Existence and uniqueness of solutions of (1.4) for a general flux f were proved by Kružkov [14]. In what follows, when talking about solutions of the differential equation $u_t + f(u)_x = 0$ in an open set, the jump condition and the entropy condition are assumed to be fulfilled along curves of discontinuity.

The Riemann problem. The main idea of the analysis of the solution of (1.2) relies heavily on classical results for the solution of the *Riemann problem*

$$\begin{aligned} &u_t + f(u)_x = 0, \quad x \in \mathbb{R},\ t > 0, \\ &u(x,0) = \begin{cases} u_l, & x < 0, \\ u_r, & x > 0, \end{cases} \end{aligned} \tag{1.6}$$

where u_l and u_r are constants. In what follows the notation $\mathrm{RP}(f; u_l, u_r)$ will be used for (1.6). The unique solution of (1.6) can be presented as follows. Assume that $u_l < u_r$. For a general flux function f, define $\breve{f}(u) = \breve{f}(u; u_l, u_r)$ on the interval $[u_l, u_r]$ by

$$\breve{f} = \sup\{h : h \text{ convex on } [u_l, u_r];\ h(v) \le f(v),\ \forall v \in [u_l, u_r]\},$$

i.e., $\breve{f}$ is the greatest convex minorant of f. The derivative $\breve{f}'$ is a continuous nondecreasing function in the interval (u_l, u_r). Let $\breve{f}'(u_l)$ be interpreted as a right derivative and $\breve{f}'(u_r)$ as a left derivative. Let h denote the inverse function of the restriction of f' to the open intervals where $\breve{f}'$ is increasing. Then the unique weak solution of (1.6) in the case $u_l < u_r$ is

$$u(x,t) = \begin{cases} u_l, & x < \breve{f}'(u_l)t, \\ h(\frac{x}{t}), & \breve{f}'(u_l)t < x < \breve{f}'(u_r)t, \\ u_r, & x > \breve{f}'(u_r)t. \end{cases} \qquad t > 0, \tag{1.7}$$

Inside the cone

$$V(f; u_l, u_r) = \{(x,t) : \breve{f}'(u_l)t \le x \le \breve{f}'(u_r)t,\ t > 0\},$$

the solution consists of rarefaction waves separated by discontinuities. For example, if f is concave and $u_l < u_r$, then the cone is merely a straight half line, i.e., a shock. The case in which $u_l > u_r$ is treated in the same way, using the least concave majorant instead. If $u_l = u_r$, then the solution is simply $\equiv u_l$ and the cone $V = \emptyset$ by definition.

2. Existence and uniqueness. In §2.1 the existence of solutions of (1.2) or (1.3) is characterized in terms of conditions on f, g, and s. The solution of (1.3) is locally governed by the equations $u_t + f(u)_x = 0$, $x > 0$, and $u_t + g(u)_x = 0$, $x < 0$, separately, away from the t-axis. To obtain a global solution, we must find boundary functions $\alpha(t)$ and $\beta(t)$ along the t-axis, which together with the initial data define solutions in $x \lessgtr 0$, respectively, such that the fitting of these two solutions defines a global solution $u(x,t)$ satisfying $u^+(t) = \alpha(t)$, $u^-(t) = \beta(t)$, and the third equation of (1.3),

$$f\big(u^+(t)\big) = g\big(u^-(t)\big) + s(t). \tag{2.1}$$

Note that $s(t)$ is continuous from the right. The key problem is that $\alpha(t)$ and $\beta(t)$ cannot be given beforehand in the general case.

The existence of solutions locally in t is proved in §2.4 by choosing allowable $\alpha(t)$ and $\beta(t)$ such that construction of solutions in $x \lessgtr 0$, respectively, by the method of characteristics is possible.

Nonunique solutions occur when more than one pair of boundary functions (α, β) is allowable. A coupling condition, Condition Γ, is introduced in §2.2 as a means to pick out a unique solution.

2.1. Definitions and necessary conditions. We start with an example showing that there may not exist any solution at all of (1.2).

Example 2.1. Consider the problem when $s \equiv$ constant, $f(u) = -u$, $g(u) = u$ and the initial data

$$u(x,0) = \begin{cases} u_l, & x < 0, \\ u_r, & x > 0. \end{cases}$$

In this case the characteristics, emanating from the x-axis and carrying the values u_l and u_r, go into the t-axis. Hence $u^-(t) = u_l$ and $u^+(t) = u_r$, $t \geq 0$. A solution of (1.2) exists if and only if $-u_r = u_l + s$, i.e., (2.1) is satisfied.

Suppose the problem (1.2) is solved up to time t, and hence $u_\pm$ are known. To characterize which $u^\pm$ are possible at time t to continue a solution, the following definitions and two lemmas are needed, in which the underlying idea comes from the knowledge of the solution of the Riemann problem (1.6).

DEFINITION 2.2. *Given $u_+, u_- \in \mathbb{R}$ and the flux functions f and g, define (see Figs. 2.1 and 2.2)*

$$\hat{f}(u;u_+) = \begin{cases} \min\limits_{v\in[u,u_+]} f(v), & u \leq u_+, \\ \max\limits_{v\in[u_+,u]} f(v), & u > u_+, \end{cases}$$

$$P(f;u_+) = \{u_+\} \cup \{u : u < u_+;\ \hat{f}(u+\varepsilon;u_+) > \hat{f}(u;u_+),\ \forall \varepsilon > 0\}$$
$$\cup \{u : u > u_+;\ \hat{f}(u-\varepsilon;u_+) < \hat{f}(u;u_+),\ \forall \varepsilon > 0\},$$

$$\tilde{P}(f;u_+) = \{u : \hat{f}(u;u_+) = f(u)\},$$

$$\check{g}(u;u_-) = \left.\begin{cases} \max\limits_{v\in[u,u_-]} g(v), & u \leq u_- \\ \min\limits_{v\in[u_-,u]} g(v), & u > u_- \end{cases}\right\} = \hat{g}(u_-;u),$$

$$N(g;u_-) = \{u_-\} \cup \{u : u < u_-;\ \check{g}(u+\varepsilon;u_-) < \check{g}(u;u_-),\ \forall \varepsilon > 0\}$$
$$\cup \{u : u > u_-;\ \check{g}(u-\varepsilon;u_-) > \check{g}(u;u_-),\ \forall \varepsilon > 0\},$$

$$\tilde{N}(g;u_-) = \{u : \check{g}(u;u_-) = g(u)\}.$$

Observe that $\hat{f}(\cdot;u_+)$ is a nondecreasing function whose graph consists of increasing parts separated by plateaus, where the function is constant. Analogously, $\check{g}(\cdot;u_-) = \hat{g}(u_-;\cdot)$ is nonincreasing with a graph consisting of decreasing parts separated by plateaus. Furthermore, $\hat{f}$ and $\check{g}$ are continuous functions in both of their variables. In what follows we shall sometimes use the shorter notation $P = P(f;u_+)$ and $N = N(g;u_-)$. The difference between P and $\tilde{P}$ is illuminated in Figs. 2.1 and 2.2 and in the following lemma. (The proof is found in [7].)

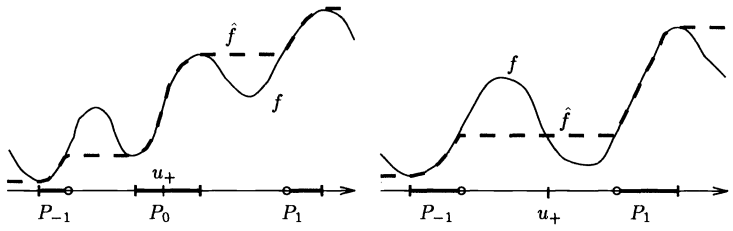


FIG. 2.1. *The graphs of f (solid) and the two main possibilities for $\hat{f}(\cdot;u_+)$ (dashed). The set $P=\bigcup_i P_i$, where $P_0=\{u_+\}$ in the right-hand plot.*

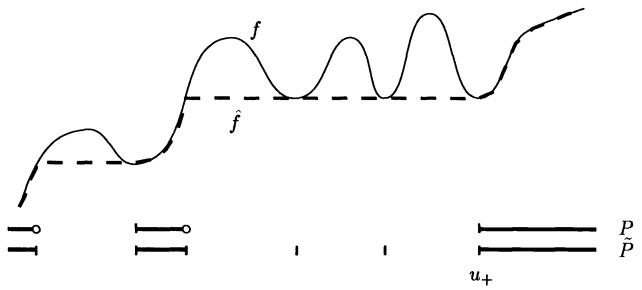


FIG. 2.2. *Examples of the sets P and $\tilde{P}$ for given u_+. Observe that $P\subseteq\tilde{P}$.*

LEMMA 2.3. *Given $u_+,u_-\in\mathbb{R}$, then*

$$P(f;u_+)=\{\alpha:\text{the solution of RP}(f;\alpha,u_+)\text{ satisfies }u^+(0)=\alpha\},$$
$$\tilde{P}(f;u_+)=\{\alpha:\text{the solution of RP}(f;\alpha,u_+)\text{ satisfies }u^-(0)=\alpha\},$$
$$N(g;u_-)=\{\beta:\text{the solution of RP}(g;u_-,\beta)\text{ satisfies }u^-(0)=\beta\},$$
$$\tilde{N}(g;u_-)=\{\beta:\text{the solution of RP}(g;u_-,\beta)\text{ satisfies }u^+(0)=\beta\}.$$

Consider a piecewise smooth solution of (1.2) in the neighbourhood of the t-axis at some time, say $t=0$. The resolution of a discontinuity approximates the solution of the corresponding Riemann problem; cf. Dafermos [5] and Chang and Hsiao [3]. Since $u_0(x)$ and $u^+(t)$ are smooth for small $x>0$ and $t>0$, respectively, the solution u of (1.2) approximates, for small $x>0$ and $t>0$, the solution of $\text{RP}\big(f;u^+(0),u_+(0)\big)$ with the cone $V\big(f;u^+(0),u_+(0)\big)\subset\{(x,t):x\geq 0,t>0\}$. An analogous reasoning holds for $x<0$. These facts yield the following lemma.

LEMMA 2.4. *If u is a piecewise smooth solution of (1.2) for $t\in[0,T)$ for some $T>0$, then*

$$u^+(t)\in\tilde{P}\big(f;u_+(t)\big),\quad t\in[0,T)$$
$$u^-(t)\in\tilde{N}\big(g;u_-(t)\big),\quad t\in[0,T).$$

DEFINITION 2.5. *Let t be fixed and $u_+,u_-\in\mathbb{R}$ be given. Define the set of intersecting ranges*

$$I(u_+,u_-,t)=\hat{f}(\mathbb{R};u_+)\cap\big(\check{g}(\mathbb{R};u_-)+s(t)\big).$$

For the projection on the u-axis of the intersection of the graphs of the functions we

define

$$\bar{U}(t) = \bar{U}(u_+, u_-, t) = \{u \in \mathbb{R} : \hat{f}(u; u_+) = \check{g}(u; u_-) + s(t)\},$$
$$\bar{u}_{\max}(t) = \sup \bar{U}(t), \quad \bar{u}_{\min}(t) = \inf \bar{U}(t).$$

Since $\hat{f}$ is nondecreasing and $\check{g}$ is nonincreasing, $\bar{U}(t)$ is an interval. When the set $\bar{U}(t)$ only consists of one point, it is denoted by $\bar{u}(t)$. Furthermore, we introduce the set of pairs

$$\Gamma(u_+, u_-, t) = \{(\alpha, \beta) \in \mathbb{R}^2 : f(\alpha) = g(\beta) + s(t) = \hat{f}(\bar{U}(t); u_+)\};$$

see Fig. 2.3.

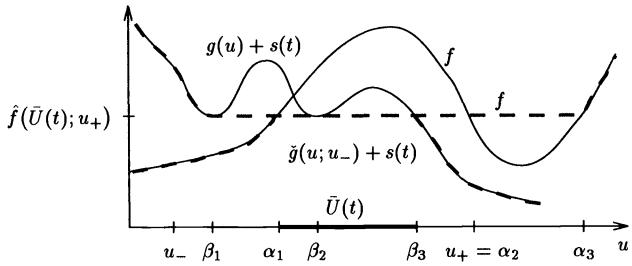


FIG. 2.3. *An example of the set $\bar{U}(t)$. The dashed line from β_1 to β_3 is a plateau of $\check{g}(\cdot; u_-) + s(t)$, and the one from α_1 to α_3 is a plateau of $\hat{f}(\cdot; u_+)$. Note that $\Gamma = \{(\alpha_i, \beta_j) : i, j = 1, 2, 3\}$.*

We say that the graphs of $\hat{f}(\cdot; u_+)$ and $\check{g}(\cdot; u_-) + s(t)$ intersect if $I(u_+, u_-, t) \neq \emptyset$. The necessary conditions on the boundary limits $u^{\pm}(t)$ can now be summarized.

THEOREM 2.6 (necessary conditions). *If u is a piecewise smooth solution of* (1.2) *for $t \in [0, T)$ for some $T > 0$, then*

$$\begin{aligned}
&(u^+(t), u^-(t)) \in \tilde{P}(f; u_+(t)) \times \tilde{N}(g; u_-(t)), && t \in [0, T),\\
&(u^+(t), u^-(t)) \in P(f; u_+(t)) \times N(g; u_-(t)), && t \in [0, T) \setminus D,\\
&f(u^+(t)) = g(u^-(t)) + s(t), && t \in [0, T),\\
&I(u^+(t), u^-(t), t) \neq \emptyset, && t \in [0, T),
\end{aligned}$$

where D is a discrete set such that

$$D \subseteq \{t_0 \in [0, T) : u_+(t) \text{ or } u_-(t) \text{ is discontinuous at } t = t_0\}.$$

Proof. The first statement is Lemma 2.4. A piecewise smooth solution satisfies $u^+(t) = u_+(t)$ for $t \in [0, T) \setminus D^+$ for some discrete set D^+. For such a t $u^+(t) \in P(f; u_+(t))$ holds by definition. Analogously, $u^-(t) \in N(g; u_-(t))$ holds for $t \in [0, T) \setminus D^-$ for some discrete set D^-. Letting $D = D^+ \cup D^-$, the second statement is proved, but this can also be true for a set D strictly contained in $D^+ \cup D^-$; see, for example, the Riemann problem with discontinuous flux function, §2.3. The third statement is (2.1). This, together with Lemma 2.4, implies $\hat{f}(u^+(t); u_+(t)) = \check{g}(u^-(t); u_-(t)) + s(t)$ for all $t \in [0, T)$. Since $\hat{f}$ is nondecreasing and $\check{g}$ is nonincreasing there must be an intersection, and the fourth statement is proved. □

2.2. The coupling condition: Condition Γ. The nonuniqueness of solutions of (1.2) is demonstrated by the following two examples.

Example 2.7. Let $s \equiv$ constant, $f(u) = u$, $g(u) = -u$, and the initial data

$$u(x,0) = \begin{cases} u_l, & x < 0, \\ u_r, & x > 0. \end{cases}$$

Independently of the values $u^{\pm}$, the characteristics always emanate from the t-axis, and hence there exist infinitely many solutions which satisfy (2.1): $u^+ = -u^- + s$. Note that $P = N = \mathbb{R}$, $\forall t \geq 0$, independently of the initial data. Two possible choices

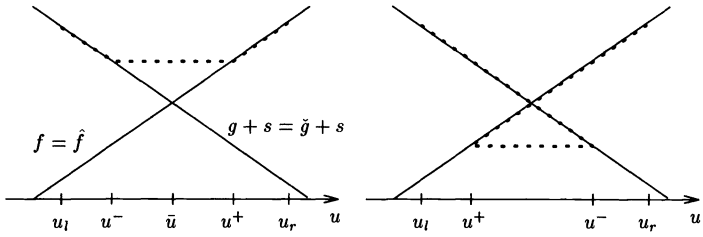


FIG. 2.4. (Example 2.7). *There are infinitely many choices of* u^- *and* u^+.

of u^+ and u^- are shown in Fig. 2.4, where the jump from u^- to u^+ corresponds to a horizontal dotted line from the graph of $g(\cdot) + s$ to the graph of f. Of all these solutions the one with $u^- = u^+ = \bar{u}$ turns out to play a distinguished role; see the numerical justification in Theorem 3.1.

Example 2.8. Let $s \equiv$ constant, $f(u) = u$, $g(u) =$ a parabola according to Figs. 2.5 and 2.6, and the initial data

$$u(x,0) = \begin{cases} u_l, & x < 0, \\ u_r, & x > 0. \end{cases}$$

Let u_1 and u_2 be the concentrations defined by Fig. 2.5, i.e., $g(u_l) + s = f(u_1) = g(u_2) + s$. The solution shown in Fig. 2.5 is in accordance with the numerical treatment

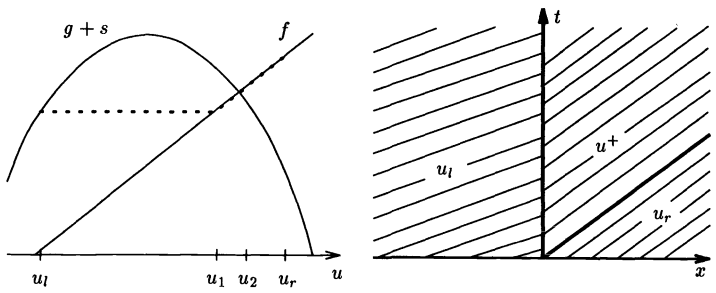


FIG. 2.5. (Example 2.8). *A solution with minimal variation. Note that* $u^- = u_l$ *and* $u^+ = u_1$. *Thin lines in the right plot are characteristics. The two dotted line segments in the left plot correspond to the two discontinuities in the right plot.*

in §3; see Theorem 3.2. Another solution is shown in Fig. 2.6.

Gimse and Risebro [11] use the condition to minimize $|u^+ - u^-|$ to obtain a unique solution of the Riemann problem with discontinuous flux function. However,

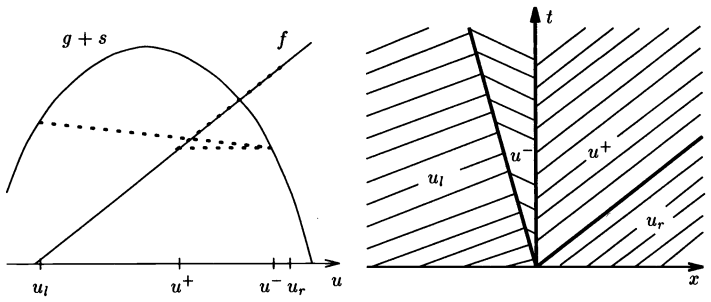


FIG. 2.6. (Example 2.8). *A solution with a smaller jump at $x = 0$ than the solution in Fig. 2.5, but with greater variation.*

Example 2.8 shows that such a jump need not exist. If we interpret the notation u^+ and u^- in [11] as "inner states" at $x = 0$ instead of as limits of a solution, then a discontinuity along the t-axis (with zero speed) would be allowed with an "inner state" on one side of the discontinuity. Then the condition to minimize $|u^+ - u^-|$ can be used as a uniqueness condition. For example, the solution in Fig. 2.5 could be seen as the limit of solutions of the type in Fig. 2.6 when the speed of the shock in $x < 0$ tends to zero from below.

Since we interpret $u^{\pm}$ as limits of a solution, the uniqueness condition presented below is based on the intersection of the graphs of $\hat{f}$ and $\check{g} + s(t)$. Some properties of a unique solution, involving the jump between u^- and u^+ and the variation of the solution, can be found in [7].

In each of the Examples 2.7 and 2.8 a specific solution can be selected by the conservative Godunov scheme of §3. Note that in Example 2.7 $\Gamma(u_+, u_-, 0) = \{(\bar{u}, \bar{u})\}$ holds; see Definition 2.5. Letting the pair of boundary functions $\big(\alpha(t), \beta(t)\big) \equiv (\bar{u}, \bar{u})$, $t \geq 0$, we then obtain the same solution that we get by the numerical treatment in §3; see Theorem 3.1. In Example 2.8 $\Gamma(u_+, u_-, 0) = \{(u_1, u_l), (u_1, u_2)\}$ holds. The solution obtained by the numerical treatment (see Theorem 3.2) coincides with the analytical solution obtained by using the boundary functions $\big(\alpha(t), \beta(t)\big) \equiv (u_1, u_l)$, $t \geq 0$. In both Examples 2.7 and 2.8 $\big(u^+(t), u^-(t)\big) \in \Gamma\big(u_+(t), u_-(t), t\big)$, $\forall t \geq 0$ holds. Motivated by this as well as by a viscous profile analysis in [8], we introduce the following coupling condition.

Condition Γ. For fixed t and given $u_+, u_- \in \mathbb{R}$, $(u^+, u^-) \in \Gamma(u_+, u_-, t)$ holds.

Since problem (1.2) or (1.3) is a generalization of (1.1), Condition Γ must be a generalization of the entropy condition.

PROPOSITION 2.9. *If $f \equiv g$ and $s \equiv 0$, then Condition Γ is equivalent to the entropy condition* (1.5).

Proof. Let $x = x(t) \in C^1$ be a discontinuity with u smooth on both sides. By a change of coordinates, under which the entropy condition is invariant, we can assume that the discontinuity has zero speed (replace x by $x - x(t)$ and $f(u)$ by $f(u) - x'(t)u$).

Assume that $u^- < u^+$ (the case $u^- > u^+$ is similar). Then

$$\begin{aligned}
\text{Condition } \Gamma &\iff (u^+, u^-) \in \Gamma(u^+, u^-, t) \\
&\iff f(u^+) = f(u^-) = \hat{f}(\bar{U}; u^+) \\
&\iff f(u^+) = f(u^-) = \hat{f}(u^-; u^+) \quad (\text{since } u^- \in \bar{U}) \\
&\iff f(u^+) = f(u^-) = \hat{f}(u^-; u^+) = \min_{u^- \le v \le u^+} f(v) \\
&\iff S(u^-, v) \ge 0 = S(u^-, u^+) \quad \forall v \text{ between } u^- \text{ and } u^+. \qquad \square
\end{aligned}$$

The following example shows (unfortunately) that there exists a time ($t = 0$ in the example) at which $(u^+, u^-) \notin \big(P(f; u_+) \times N(g; u_-)\big)$, i.e., the set D in Theorem 2.6 can be nonempty. It also provides another example of a solution satisfying Condition Γ.

Example 2.10. Let $f(u) = u$, $g(u) =$ a parabola according to Fig. 2.7, $s(t) = t$, and the initial data

$$u(x, 0) = \begin{cases} u_l, & x < 0, \\ u_r, & x > 0. \end{cases}$$

where u_r is arbitrary and u_l satisfies $f\big(\bar{u}(0)\big) = g\big(\bar{u}(0)\big) = g(u_l)$ according to the figure. Since $s(t)$ is increasing, we can let $u^i(t)$ be the smooth increasing function that is the unique intersection of the graphs of $f(\cdot)$ and $g(\cdot) + s(t)$ with $u^i(0) = \bar{u}(0)$. The solution shown in Fig. 2.7 satisfies Condition Γ, and $u^+(t) = u^-(t) = u^i(t)$ for $t \ge 0$ and, in particular, $u^-(0) = \bar{u}(0) \in \tilde{N}\big(u_-(0)\big) \setminus N\big(u_-(0)\big)$, where $u_-(0) = u_l$. Note that the set $\Gamma(u_r, u_l, 0) = \big\{\big(\bar{u}(0), u_l\big), \big(\bar{u}(0), \bar{u}(0)\big)\big\}$ consists of two pairs, but

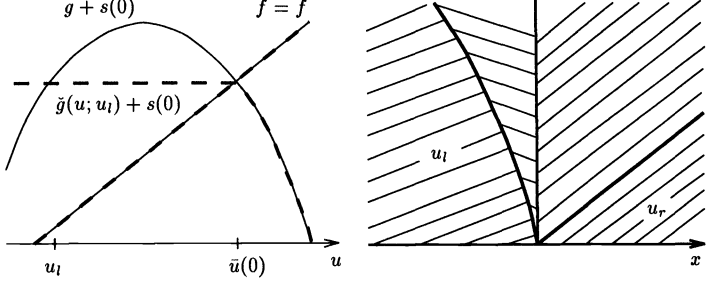


FIG. 2.7. (Example 2.10). *A shock moves to the left, separating the value u_l from the values $u^-(t) = u^i(t)$, $t \ge 0$, on the characteristics coming from the t-axis. In $x > 0$ the characteristics emanating from the t-axis carry the values $u^+(t) = u^i(t)$, $t \ge 0$.*

since $s(t)$ is increasing, only the pair $\big(\bar{u}(0), \bar{u}(0)\big)$ serves as the initial condition for the pair of boundary functions $\big(\alpha(t), \beta(t)\big)$. If $s(t)$ were decreasing, only the other pair of $\Gamma(u_r, u_l, 0)$ would be possible; cf. Example 2.8. This will follow from Theorem 2.19.

2.3. The Riemann problem with discontinuous flux function. Let the source function s be independent of time. It is no restriction to let $s = 0$. Using the same initial data as in the Riemann problem (1.6), (1.3) becomes the *Riemann*

problem with discontinuous flux function

$$
\begin{aligned}
&u_t + f(u)_x = 0, && x > 0,\ t > 0,\\
&u_t + g(u)_x = 0, && x < 0,\ t > 0,\\
&f\big(u^+(t)\big) = g\big(u^-(t)\big), && t > 0,\\
&u(x,0) = \begin{cases} u_l, & x < 0,\\ u_r, & x > 0. \end{cases}
\end{aligned}
\tag{2.2}
$$

This problem is treated by Gimse and Risebro [11]; cf. the discussion after Example 2.8 above. A solution of this problem can be constructed by fitting two "Riemann cones" with $(u^+, u^-) \in P \times N$ according to Lemma 2.3. The following proposition introduces a function $c(u_+, u_-, t) = (u^+, u^-) \in (P \times N) \cap \Gamma$, which will yield the correct boundary values. Furthermore, if u is a solution of (1.2) satisfying Condition Γ, then Theorem 2.6 says that $(u^+, u^-) \in (P \times N) \cap \Gamma$ for all t outside a discrete set. Therefore, the function c will also be used in the construction of solutions in the general case; see §2.4.

PROPOSITION 2.11. *Let t be fixed and $u_+, u_- \in \mathbb{R}$ be given. If $I(u_+, u_-, t) \neq \emptyset$, then the set $\big(P(f; u_+) \times N(g; u_-)\big) \cap \Gamma(u_+, u_-, t)$ consists of exactly one point, and hence a function c is well defined by*

$$c(u_+, u_-, t) = (u^+, u^-) \in (P \times N) \cap \Gamma.$$

Proof. $I \neq \emptyset$ implies $\bar{U} \neq \emptyset$. Put $\gamma = \hat{f}(\bar{U}; u_+)$. Since the restrictions $f|_P$ and $g|_N$ are injective,

$$(u^+, u^-) \in (P \times N) \cap \Gamma \quad \Longleftrightarrow \quad f|_P\,(u^+) = g|_N\,(u^-) + s(t) = \gamma$$

uniquely determines u^+ and u^-. □

We shall now describe the function c by considering all the cases that may occur depending on the set $\bar{U}$.

Case 1. $\bar{u} \in N \cap P$; see Fig. 2.8 and Example 2.7. Application of the function c yields $u^+ = u^- = \bar{u}$.

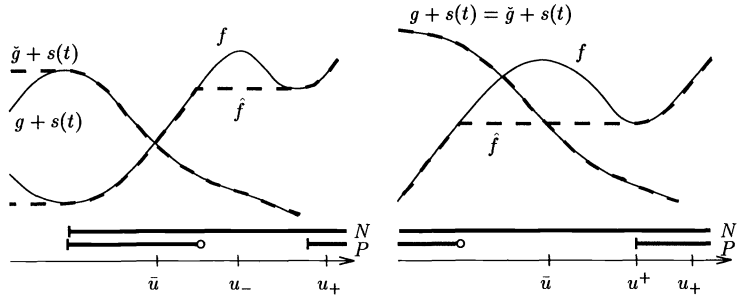


FIG. 2.8. *Examples of Cases* 1 (*left*) *and* 2a (*right*). *The sets N and P are indicated above the u-axis.*

Case 2a. $\bar{u} \in N \setminus P$; see Fig. 2.8. The function c yields

$$
\begin{aligned}
u^+ &= \begin{cases} \max\big(P \cap (-\infty, \bar{u})\big), & u_+ < \bar{u},\\ \min\big(P \cap (\bar{u}, \infty)\big), & u_+ > \bar{u}, \end{cases}\\
u^- &= \bar{u}.
\end{aligned}
$$

Case 2b. $\bar{u} \in P \setminus N$ (symmetric to Case 2a); cf. Example 2.8. The function c yields

$$u^+ = \bar{u},$$
$$u^- = \begin{cases} \max\big(N \cap (-\infty, \bar{u})\big), & u_- < \bar{u}, \\ \min\big(N \cap (\bar{u}, \infty)\big), & u_- > \bar{u}. \end{cases}$$

Case 3. $\bar{U}$ is infinite or $\bar{u} \notin (P \cup N)$; see Fig. 2.9. The function c yields

$$u^+ = \begin{cases} \max\big(P \cap (-\infty, \bar{u}_{\min}]\big), & u_+ < \bar{u}_{\min}, \\ u_+, & \bar{u}_{\min} \le u_+ \le \bar{u}_{\max}, \\ \min\big(P \cap [\bar{u}_{\max}, \infty)\big), & u_+ > \bar{u}_{\max}, \end{cases}$$
$$u^- = \begin{cases} \max\big(N \cap (-\infty, \bar{u}_{\min}]\big), & u_- < \bar{u}_{\min}, \\ u_-, & \bar{u}_{\min} \le u_- \le \bar{u}_{\max}, \\ \min\big(N \cap [\bar{u}_{\max}, \infty)\big), & u_- > \bar{u}_{\max}. \end{cases}$$

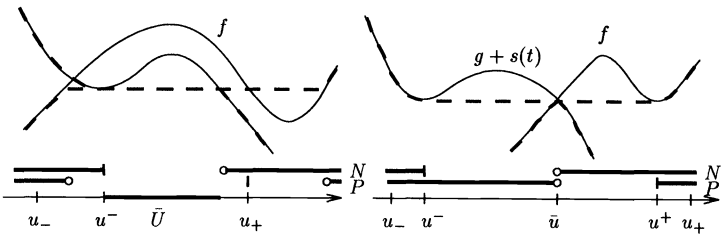


FIG. 2.9. *Examples of Case* 3: $\bar{U}$ *is infinite (left) and* $\bar{u} \notin (P \cup N)$ *(right). In the left figure* $u^+ = u_+$ *holds.*

THEOREM 2.12. *If* $I(u_r, u_l, 0) \neq \emptyset$, *then there exists a unique solution* $u \in \Sigma$ *of* (2.2) *satisfying Condition* Γ *for* $t \ge 0$. *The solution is of the form* $u(\frac{x}{t})$ *and the constant boundary values are given by* $(u^+, u^-) = c(u_r, u_l, 0)$.

Proof. Let $(\alpha_0, \beta_0) = c(u_r, u_l, 0)$. By Lemma 2.3 there exists a solution, say $v(\frac{x}{t})$, of $\mathrm{RP}(f; \alpha_0, u_r)$ such that the cone $V(f; \alpha_0, u_r)$ is entirely contained in $x \ge 0$, $t > 0$, and $v(0+) = \alpha_0$. Analogously, there is a solution $w(\frac{x}{t})$ of $\mathrm{RP}(g; u_l, \beta_0)$ with a cone entirely in $x \le 0$, $t > 0$, and with $w(0-) = \beta_0$. Since $f(\alpha_0) = g(\beta_0)$ by the definition of c, a function solving the problem is

$$u\left(\frac{x}{t}\right) = \begin{cases} v(\frac{x}{t}), & x > 0, \\ w(\frac{x}{t}), & x < 0, \qquad t > 0. \end{cases}$$

To prove the uniqueness let $\tilde{u}(x,t) \in \Sigma$ be any solution of (2.2) that satisfies Condition Γ $\forall t \ge 0$. First we show that $\big(\tilde{u}^+(t), \tilde{u}^-(t)\big) = (\alpha_0, \beta_0)$ for small $t > 0$. Assume that $\tilde{u}^+(t)$ is nonconstant for small $t > 0$. Since $\tilde{u}^+(t)$ is smooth and monotone for small $t > 0$ and f nonconstant on every open interval, two cases may appear:

1. $f'\big(\tilde{u}^+(t)\big) < 0$ for small $t > 0$. Then the characteristics to the right of the t-axis have negative speed and must therefore come from the positive x-axis. Hence $\tilde{u}^+(t) \equiv u_r$ for small $t > 0$, which is a contradiction.

2. $f'\big(\tilde{u}^+(t)\big) > 0$ for small $t > 0$. Since $\tilde{u}^+(t)$ is nonconstant and g is nonconstant on every open interval, the relation (2.1), $f\big(\tilde{u}^+(t)\big) = g\big(\tilde{u}^-(t)\big)$, implies that

$\tilde{u}^-(t)$ is nonconstant for small $t > 0$. Thus $g\big(\tilde{u}^-(t)\big) = \check{g}\big(\tilde{u}^-(t); \tilde{u}_-(t)\big)$ is also nonconstant and Condition Γ implies that this occurs only if $g'\big(\tilde{u}^-(t)\big) > 0$ for small $t > 0$. Then the characteristics to the left of the t-axis have positive speed and must come from the negative x-axis carrying the constant value $\tilde{u}^-(t) \equiv u_l$, which is also a contradiction.
Thus $\tilde{u}^+(t) \equiv$ constant for small $t > 0$ must hold and, then,

$$\tilde{u}(x,t) = \tilde{u}^{\mathrm{RP}}\left(\frac{x}{t}\right) \quad \text{for } x > 0 \text{ and small } t > 0, \tag{2.3}$$

where $\tilde{u}^{\mathrm{RP}}$ is the solution of $\mathrm{RP}\big(f; \tilde{u}^+(0), u_r\big)$. Theorem 2.6 and Condition Γ say that $\big(\tilde{u}^+(0), \tilde{u}^-(0)\big) \in \big(\tilde{P}(u_r) \times \tilde{N}(u_l)\big) \cap \Gamma(u_r, u_l, 0)$, which implies that $\tilde{u}^+(0) = \tilde{u}^{\mathrm{RP}}(0+) = \alpha_0$, and hence $\tilde{u}^+(t) \equiv \alpha_0$ for small $t > 0$. Analogously, we conclude that $\tilde{u}^-(t) \equiv \beta_0$ for small $t > 0$. The only possibility left for $\tilde{u}(x,t)$ to differ from $u(\frac{x}{t})$ is that $\tilde{u}^+(t) \neq \alpha_0$ and/or $\tilde{u}^-(t) \neq \beta_0$ for $t > t_0 \in (0, \infty)$. Then the "new initial data" are $\tilde{u}(x, t_0) = u(x/t_0) = v(x/t_0)$, $x > 0$. Either $\tilde{u}(x, t_0) \equiv u_r$, $x > 0$; then $\alpha_0 = u_r$, and the reasoning above (at $t = 0$) gives $\tilde{u}^+(t) \equiv u_r = \alpha_0 = u^+(t)$ for small $t - t_0 > 0$. Otherwise $v(\frac{x}{t})$ consists of a cone $V(f; \alpha_0, u_r)$ entirely contained in $x \geq 0$, so that $f'\big(v(\frac{x}{t})\big) > 0$ for small $x > 0$. Almost the same reasoning as above (at $t = 0$) can be used: If $\tilde{u}^+(t)$ is nonconstant for small $t - t_0$, then only item 2 is possible and the same reasoning holds and gives a contradiction. Thus $\tilde{u}^+(t) \equiv$ constant for $0 < t < t_0 + \varepsilon$ for some $\varepsilon > 0$, and this implies that (2.3) holds for $0 < t < t_0 + \varepsilon$, so the limit value is again $\tilde{u}^+(t) \equiv \alpha_0$ for $0 < t < t_0 + \varepsilon$, and we have proved that $\tilde{u}^+(t) \equiv u^+(t)$ for $0 < t < t_0 + \varepsilon$. Analogously, $\tilde{u}^-(t) \equiv u^-(t)$ holds for small $t - t_0 > 0$ and hence $\tilde{u}(x,t) \equiv u(\frac{x}{t})$ for small $t - t_0 > 0$, which is a contradiction. □

2.4. Construction of solutions in the general case. In this section the existence of a solution of (1.2) locally in t is proved by construction. When constructing solutions of the simpler problem (1.1), certain assumptions have to be laid on the initial data; see Ballou [1] and Cheng [4]. To construct a solution of (1.2) we must define boundary functions, $\alpha(t)$ and $\beta(t)$ with the same regularity that we require of the initial data u_0, i.e., piecewise smoothness and piecewise monotonicity. However, since the main problem of the construction is to define these boundary functions, we must impose restrictions on u_0, f, g, and s to ensure that $\alpha(t)$ and $\beta(t)$ become piecewise smooth and piecewise monotone. Since the behaviour of a solution changes abruptly when a discontinuity reaches the t-axis, it is natural to formulate conditions for existence locally in t. In Definition 2.14 restrictions on u_0, f, g, and s are given, defining what we call a *regular* problem. If problem (1.2) is regular, then we show how a solution $u \in \Sigma$ satisfying Condition Γ can be constructed by the method of characteristics for $0 \leq t < \varepsilon$ for some $\varepsilon > 0$. Then, if $u(x, \varepsilon - 0)$ serves as initial data for a new regular problem starting at $t = \varepsilon$, the solution can be continued. We comment further on Definition 2.14 at the end of this section.

Below we shall present a "procedure of construction" of a solution and postpone, until Theorem 2.17, the proof that it works and that the solution satisfies Condition Γ, as well as the fact that it belongs to the class Σ. Also, the proof of Theorem 2.17 will clarify the steps of the procedure. The idea is the following. From Theorem 2.6 we know that a piecewise smooth solution satisfying Condition Γ fulfils $(u^+, u^-) = c(u_+, u_-, t)$ for all t outside a discrete set. In contrast to the simpler problem in the previous section, the function c defined in Proposition 2.11 will be used twice to define the boundary functions $\alpha(t)$ and $\beta(t)$ in the general case. This is because of the dependence of $u_0(x)$ on x and the dependence of $s(t)$ on t. At $t = 0$ the

function c is used for the first time to define two constants a and b, which are used in two auxiliary problems. These problems produce two functions $\tilde{v}^+(t)$ and $\tilde{w}^-(t)$, on which the function c is applied again to define, finally, $\alpha(t)$ and $\beta(t)$.

It is convenient to divide the initial data

$$u(x,0) = \begin{cases} u_l(x), & x < 0, \\ u_r(x), & x > 0, \end{cases}$$

and define $u_l(0) \equiv u_l(0-)$ and $u_r(0) \equiv u_r(0+)$.

PROCEDURE OF CONSTRUCTION.

1. Let $(a,b) = c\big(u_r(0), u_l(0), 0\big)$.
2. Solve the initial value problems

$$\begin{aligned} &\tilde{v}_t + f(\tilde{v})_x = 0, && \tilde{w}_t + g(\tilde{w})_x = 0, \\ &\tilde{v}(x,0) = \begin{cases} a, & x < 0, \\ u_r(x), & x > 0, \end{cases} \quad \text{and} && \tilde{w}(x,0) = \begin{cases} u_l(x), & x < 0, \\ b, & x > 0, \end{cases} \end{aligned} \tag{2.4}$$

and compute $\tilde{v}^+(t)$ and $\tilde{w}^-(t)$ for $t \geq 0$.

3. Let

$$T = \sup\big\{t_1 : I\big(\tilde{v}^+(t), \tilde{w}^-(t), t\big) \neq \emptyset, \forall t \in [0, t_1]\big\} \tag{2.5}$$

and define

$$\begin{aligned} &\big(\alpha(t), \beta(t)\big) = c\big(\tilde{v}^+(t), \tilde{w}^-(t), t\big), \quad 0 < t < T, \\ &\big(\alpha(0), \beta(0)\big) = \big(\alpha(0+), \beta(0+)\big). \end{aligned} \tag{2.6}$$

4. Solve the initial boundary value problems with $\varepsilon \leq T$ as large as possible, i.e., ε is the first time when $\alpha(t)$ or $\beta(t)$ is discontinuous or a discontinuity reaches the t-axis:

$$\begin{aligned} &v_t + f(v)_x = 0, && x > 0,\ 0 < t < \varepsilon, \\ &v(x,0) = u_r(x), && x > 0, \\ &v^+(t) = \alpha(t), && 0 \leq t < \varepsilon, \end{aligned} \tag{2.7}$$

and

$$\begin{aligned} &w_t + g(w)_x = 0, && x < 0,\ 0 < t < \varepsilon, \\ &w(x,0) = u_l(x), && x < 0, \\ &w^-(t) = \beta(t), && 0 \leq t < \varepsilon. \end{aligned} \tag{2.8}$$

5. Let

$$u(x,t) = \begin{cases} v(x,t), & x > 0, \\ w(x,t), & x < 0, \end{cases} \qquad 0 \leq t < \varepsilon. \tag{2.9}$$

In step 3 we define $\alpha(t)$ and $\beta(t)$ as continuous from the right, since they should eventually satisfy $\alpha(t) = u^+(t)$ and $\beta(t) = u^-(t)$.

Example 2.13. If we apply the procedure to the problem in Example 2.10, we obtain (1) $(a,b) = \big(\bar{u}(0), u_l\big)$, (2) $\tilde{v}^+(t) \equiv \bar{u}(0)$ and $\tilde{w}^-(t) \equiv u_l$, and (3) $\alpha(t) = \beta(t) = u^i(t)$, $t \geq 0$. Note that in this example $\big(\alpha(0), \beta(0)\big) \neq c\big(\tilde{v}^+(0), \tilde{w}^-(0), 0\big)$, which motivates definition (2.6) at $t = 0$. Furthermore, we can let $\varepsilon = \infty$ in 4 and the solution $u(x,t)$ in 5 is shown in Fig. 2.7.

DEFINITION 2.14. *Problem* (1.2) *is said to be* regular *if the following hold:*

1. *The solutions* $\tilde{v}$ *and* $\tilde{w}$ *of the initial value problems* (2.4) *belong to* Σ *for small* $t > 0$.

2. *The function*

$$\eta(t) \equiv f\big(\tilde{v}^+(t)\big) - g\big(\tilde{w}^-(t)\big) - s(t) \tag{2.10}$$

is either $>$, $<$, *or* $\equiv 0$ *on some interval* $0 < t < \delta$.

3. *If* $u^i(t)$ *is a unique intersection of* $f(\cdot)$ *and* $g(\cdot) + s(t)$ *for small* $t > 0$ *with* $u^i(0) = \bar{u}(0)$, $\bar{u}_{\min}(0)$ *or* $\bar{u}_{\max}(0)$, *then the functions*

$$f\big(u^i(t)\big) - f\big(\tilde{v}^+(t)\big),$$
$$g\big(u^i(t)\big) - g\big(\tilde{w}^-(t)\big)$$

are either $>$, $<$, *or* $\equiv 0$ *on some interval* $0 < t < \delta$.

For item 1, note that $\tilde{v}$ and $\tilde{w}$ are generically piecewise smooth in the case of one inflection point of f and g, respectively; see Dafermos [5]. This is the case in the applications to sedimentation and two-phase flow; see Fig. 1.2. Furthermore, the solutions $\tilde{v}$ and $\tilde{w}$ near the origin are solutions of perturbed Riemann problems with the initial data monotone on each side of $x = 0$. Considering all cases of flux functions (see Chang and Hsiao [3]) it follows that $\tilde{v}^+(t)$ and $\tilde{w}^-(t)$ are monotone for small $t \geq 0$.

Before stating the existence theorem, we need two lemmas on some properties of the function c. The first says, among other things, what happens when applying c twice.

LEMMA 2.15. *Let* $\mathrm{ch}(A, B)$ *denote the convex hull of* $A \cup B$, *where* A *and* B *are points or intervals of* $\mathbb{R}$. *Given a fixed* $t \geq 0$ *and* $u_+, u_- \in \mathbb{R}$, *let* $(\alpha_0, \beta_0) = c(u_+, u_-, t)$ *and* $\bar{U}_0 = \bar{U}(u_+, u_-, t)$. *Then*

$$\begin{cases} \alpha \in \mathrm{ch}(u_+, \alpha_0) \\ \beta \in \mathrm{ch}(u_-, \beta_0) \end{cases} \quad \Longrightarrow \quad \begin{cases} \hat{f}(u;\alpha) = \hat{f}(u;u_+) & \forall u \in \mathrm{ch}(\bar{U}_0, \alpha_0), \\ \check{g}(u;\beta) = \check{g}(u;u_-) & \forall u \in \mathrm{ch}(\bar{U}_0, \beta_0), \\ \bar{U}(\alpha, \beta, t) = \bar{U}_0, \\ c(\alpha, \beta, t) = (\alpha_0, \beta_0). \end{cases}$$

Proof. Let $\bar{U}_0 = \bar{U}(u_+, u_-, t)$. From Cases 1–3 after Proposition 2.11 it follows that either $\alpha_0 = u_+$ or α_0 lies closer to $\bar{U}_0$ than u_+ and on the same side of $\bar{U}_0$ as u_+. Analogously, either $\beta_0 = u_-$ or β_0 lies closer to $\bar{U}_0$ than u_- and on the same side of $\bar{U}_0$ as u_-. Then Definition 2.2 of $\hat{f}$, P, etc. implies the statements in turn. □

LEMMA 2.16. *Let* $\bar{U}(0) = \bar{U}\big(u_r(0), u_l(0), 0\big)$. *The solutions* $\tilde{v}$ *and* $\tilde{w}$ *of* (2.4) *satisfy*

$$\bar{U}\big(\tilde{v}^+(0), \tilde{w}^-(0), 0\big) = \bar{U}(0),$$
$$\hat{f}\big(\bar{U}(0); \tilde{v}^+(0)\big) = \hat{f}\big(\bar{U}(0); u_r(0)\big),$$
$$\check{g}\big(\bar{U}(0); \tilde{w}^-(0)\big) = \check{g}\big(\bar{U}(0); u_l(0)\big).$$

Proof. The resolution of the discontinuity between a and $u_r(0)$ is approximated by the solution of $\mathrm{RP}\big(f; a, u_r(0)\big)$; cf. the discussion preceding Lemma 2.4. This implies $\tilde{v}^+(0) \in \mathrm{ch}\big(a, u_r(0)\big)$ and $\tilde{w}^-(0) \in \mathrm{ch}\big(b, u_l(0)\big)$. The statements are implied by Lemma 2.15. □

THEOREM 2.17 (existence). *If* (1.2) *is a regular problem and*

$$I(\alpha,\beta,t)\neq\emptyset \quad \forall(\alpha,\beta,t)\in\mathbb{R}\times\mathbb{R}\times[0,T) \text{ for some } T>0, \tag{2.11}$$

then there exists a solution $u\in\Sigma$ *satisfying Condition* Γ *for* $t\in[0,\varepsilon)$ *for some* $\varepsilon\in(0,T]$.

Remark. (2.11) can be replaced by the weaker conditions $I\big(u_r(0),u_l(0),0\big)\neq\emptyset$ and $T>0$ with T defined by (2.5).

Proof. Carry out steps 1–3 in the procedure of construction. (2.6) alone gives $f\big(\alpha(t)\big)=g\big(\beta(t)\big)+s(t)$ for $t\geq 0$ and together with Lemma 2.15 gives

$$\big(\alpha(t),\beta(t)\big)=c\big(\alpha(t),\beta(t),t\big),\quad t>0. \tag{2.12}$$

In particular, (2.12) means $\big(\alpha(t),\beta(t)\big)\in\Gamma\big(\alpha(t),\beta(t),t\big)$, $t>0$. It remains to verify Condition Γ at $t=0$. The continuity of $\hat f$ implies that $\hat f\left(\bar U\big(\tilde v^+(t),\tilde w^-(t),t\big);\tilde v^+(t)\right)$ is a continuous function of t for small $t\geq 0$. Using (2.6) and the continuity of $\hat f$ and $\check g$ and letting $t\to 0+$, we get

$$f\big(\alpha(0)\big)=g\big(\beta(0)\big)+s(0)=\hat f\left(\bar U\big(\tilde v^+(0),\tilde w^-(0),0\big);\tilde v^+(0)\right).$$

This, together with Lemma 2.16, gives $\big(\alpha(0),\beta(0)\big)\in\Gamma\big(u_r(0),u_l(0),0\big)$.

Below it will be shown that $\alpha(t)$ and $\beta(t)$ are smooth and monotone for $0<t<\varepsilon$ for some $\varepsilon>0$ (we can assume that $s(t)$ is continuous in this interval). Then (2.12) and Lemma 2.3 ensure that the method of characteristics can be applied to construct solutions v and w in the strip $0\leq t<\varepsilon$ of the initial boundary value problems (2.7) and (2.8). Then $u(x,t)$ in (2.9) is a solution in Σ of (1.2) for $0\leq t<\varepsilon$. Near the origin this solution is smooth except along the t-axis and along possible discontinuities emanating from the origin going into $x>0$ and $x<0$, respectively.

By the assumption that $u(x,0)$ is piecewise smooth and piecewise monotone there exists a $\delta>0$ such that one of the following alternatives holds:

I. $f'\big(u_r(x)\big)\geq 0$ for $0<x<\delta$ and $g'\big(u_l(x)\big)\leq 0$ for $-\delta<x<0$. Then $\big(\tilde v^+(t),\tilde w^-(t)\big)\equiv(a,b)$ holds, which implies that $\hat f\big(\cdot;\tilde v^+(t)\big)\equiv\hat f(\cdot;a)$ and $\check g\big(\cdot;\tilde w^-(t)\big)\equiv\check g(\cdot;b)$ are independent of time. Since by assumption s is monotone for small $t\geq 0$, there exists an $\varepsilon_1>0$ such that one of the following cases occurs:

A. $s(t)\equiv s(0)$, $0<t<\varepsilon_1$. Then α and β defined by (2.6) are constants. Define v and w by (2.7) and (2.8). Let $\varepsilon\in(0,\varepsilon_1]$ be the first time a discontinuity crosses the t-axis. Then (2.9) defines a solution for $0<t<\varepsilon$.

B. $s(t)\neq s(0)$, $0<t<\varepsilon_1$. Then (2.6) implies that α and β satisfy

$$\hat f\big(\alpha(t);a\big)=\check g\big(\beta(t);b\big)+s(t),\quad t>0. \tag{2.13}$$

The graph of the nondecreasing function $\hat f$ consists of increasing parts separated by plateaus (where $\hat f\equiv$ constant) and, analogously, the graph of $\check g$ consists of decreasing parts separated by plateaus. Define the set $\bar U(t)=\bar U(a,b,t)$, $t\geq 0$. Three cases may occur:

(i) There is a unique intersection at $\bar u(0)$ with $f'(u)=\hat f'(u)>0$ and $g'(u)=\check g'(u)<0$ $\forall u$ $(\neq\bar u(0))$ in a neighbourhood of $\bar u(0)$. The parenthesis in the previous sentence applies if $\bar u(0)$ happens to be

an inflection point. Condition Γ implies that $\alpha(0) = \beta(0) = \bar{u}(0)$ and (2.13) reduces to

$$f\big(\bar{u}(t)\big) = g\big(\bar{u}(t)\big) + s(t),$$

which defines $\alpha(t) = \beta(t) = \bar{u}(t) \in C^1(0, \varepsilon_2)$ for some $\varepsilon_2 \in (0, \varepsilon_1]$ by the implicit function theorem. The assumptions on $s(t)$ and the fact that f is increasing and g is decreasing in a neighbourhood of $\bar{u}(0)$ yield that

$$\bar{u}'(t) = \frac{s'(t)}{f'\big(\bar{u}(t)\big) - g'\big(\bar{u}(t)\big)}$$

is either positive, negative, or zero on $0 < t < \varepsilon_3$ for some $\varepsilon_3 \in (0, \varepsilon_2]$. Hence $\alpha(t) = \beta(t) = \bar{u}(t)$ is monotone with bounded derivative for $t \in (0, \varepsilon_3)$. Let (2.7) and (2.8) define solutions v and w and let $\varepsilon \in (0, \varepsilon_3]$ be the first time a discontinuity enters the t-axis. Then (2.9) defines a solution for $0 < t < \varepsilon$.

(ii) There is a unique intersection at $\bar{u}(0)$ which separates a plateau and a strictly monotone part of $\hat{f}$ or $\check{g} + s(0)$. Since $s(t)$ is either increasing or decreasing for small $t > 0$, there is either an intersection as in (i) or an intersection with exactly one plateau involved for small $t > 0$. In the latter case either α or β is constant and the other is defined by (2.13) and is smooth with bounded derivative and monotone for $0 < t < \varepsilon_2 \in (0, \varepsilon_1]$ by the implicit function theorem and the assumptions on $s(t)$. Let (2.7) and (2.8) define solutions v and w and let $\varepsilon \in (0, \varepsilon_2]$ be the first time a discontinuity enters the t-axis. Then (2.9) defines a solution for $0 < t < \varepsilon$.

(iii) $\bar{U}(0)$ is infinite, i.e., a plateau of $\hat{f}$ coincides with a plateau of $\check{g} + s(0)$ at $t = 0$. Assumption 2 of Definition 2.14 implies that the plateaus separate immediately and we get a unique intersection as in (i).

II. $f'\big(u_r(x)\big) \geq 0$ for $0 < x < \delta$ and $g'\big(u_l(x)\big) > 0$ for $-\delta < x < 0$. Then $\tilde{v}^+(t) \equiv a$ and hence $\hat{f}$ is independent of time for small $t > 0$. $\tilde{w}^-(t)$ is defined by the characteristics from the negative x-axis carrying the values $u_l(x)$. Then we say that one plateau of the graph of $\check{g}\big(\cdot; \tilde{w}^-(t)\big)$ is *moving* and we denote the set of the corresponding u-values by

$$M(t) = \big\{u : \check{g}\big(u; \tilde{w}^-(t)\big) = g\big(\tilde{w}^-(t)\big)\big\}, \quad t > 0;$$

see Fig. 2.10. The other plateaus are called *fixed*. Also define $\bar{U}(t) = \bar{U}\big(a, \tilde{w}^-(t), t\big)$, $t \geq 0$. A solution can be constructed as in case I with some modifications depending on the moving plateau. Instead of making a division depending on $s(t)$ (IA and IB), we must consider the sign of (2.10); $\eta(t) = f(a) - g\big(\tilde{w}^-(t)\big) - s(t)$, which is either positive, negative, or zero for small $t > 0$ by the regularity assumption (item 2 of Definition 2.14). Then, for example, $\eta(t) \equiv 0$ for small $t > 0$ means that the moving plateau lies on the fixed value $f(a)$, but with one or both end points moving smoothly and monotonically (by the implicit function theorem and the assumptions on $s(t)$), and hence (2.6) will yield smooth and monotone functions $\alpha(t)$ and $\beta(t)$ for small $t > 0$ as in case I. These functions will have bounded derivatives except possibly at $t = 0$.

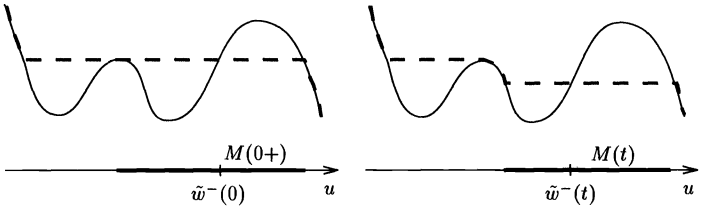


FIG. 2.10. *The moving plateau at $t = 0$ (left) and at some $t > 0$ (right) in a case when $\tilde{w}^-(t)$ is decreasing.*

One new complication arises when there is a unique intersection at $\bar{u}(0)$, which also is an end point of a moving plateau. Let us study the case when $\bar{u}(0) = \min M(0+)$ in Fig. 2.10 and $\tilde{w}^-(t)$ is decreasing. If $s(t)$ is decreasing for small $t > 0$, then there is a unique intersection of the graph of f and the fixed plateau of $\check{g}$ in Fig. 2.10, and a solution is defined as in IB(ii). If $s(t) \geq s(0)$ for small $t > 0$, then there is a unique intersection $u^i(t)$ of the graphs of $f(\cdot)$ and $g(\cdot) + s(t)$ for small $t > 0$ with $u^i(0) = \bar{u}(0)$. Then item 3 of the regularity assumption gives the fact that $g\big(u^i(t)\big) - g\big(\tilde{w}^-(t)\big)$ is either negative or nonnegative for small $t > 0$, i.e., either the graph of f intersects the plateau or the decreasing part of $\check{g}\big(\cdot; \tilde{w}^-(t)\big)$ for small $t > 0$. Hence a solution is defined either as in IB(i) or IB(ii).

III. $f'\big(u_r(x)\big) < 0$ for $0 < x < \delta$ and $g'\big(u_l(x)\big) \leq 0$ for $-\delta < x < 0$. This case is symmetrical to the previous one.

IV. $f'\big(u_r(x)\big) < 0$ for $0 < x < \delta$ and $g'\big(u_l(x)\big) > 0$ for $-\delta < x < 0$. In this case both $\hat{f}$ and $\check{g}$ have moving plateaus. Because of assumptions 2 and 3 of Definition 2.14, a solution can be constructed with an extension of cases II and III similar to the extension from case I to case II.

Finally, note that the constructed solution belongs to Σ. □

The technical reason for Definition 2.14 is that we want to avoid plateaus (of $\hat{f}$ or $\check{g}$) oscillating with unbounded frequencies. The restrictions thus mean that a plateau either stays fixed or moves monotonically (for small $t > 0$) away from, for example, another plateau. In each case only one particular pair of the set Γ will be possible as initial value for the pair of boundary functions (α, β); see §2.5. However, even though three different solutions appear in these cases of movements of a plateau, these three solutions approximate each other for small $t > 0$. This is true because, regardless of what pair of Γ is chosen, a possible discontinuity between u_- and u^- (or u_+ and u^+) will have zero speed for $t = 0+$; cf. the discussion after Example 2.8. This indicates that oscillations of bounded variation would not cause any trouble and it seems plausible that the regularity assumption in Definition 2.14 could be relaxed considerably. However, for the applications we have in mind it is very unlikely that one should find a situation in which the problem is not regular, and hence the procedure of construction could be used repeatedly to obtain a global solution.

Note that letting either $u_l(x)$ and $u_r(x)$ be constant or $s(t)$ be constant does not simplify the construction of a solution and the proof of existence in this section. If all three functions are constant, we have the Riemann problem with discontinuous flux function (2.2).

2.5. Proof of uniqueness. In this section we shall outline how to prove that the solution constructed by the method in §2.4 is the only one in the class Σ that

satisfies Condition Γ. The proofs of Theorems 2.18, 2.19, and 2.20 below provide many examples of the construction of a solution. The same notation as in §2.4 will be used.

Consider the three cases after Proposition 2.11. When there is a unique intersection as in Case 1 (in an open time interval) it is easy to construct a solution satisfying Condition Γ. It will simply satisfy $u^+(t) = u^-(t) = \bar{u}(t)$. This solution is trivially unique since the set $\Gamma\big(u_+(t), u_-(t), t\big) = \Gamma\big(\bar{u}(t), \bar{u}(t), t\big) = \big\{(\bar{u}(t), \bar{u}(t))\big\}$ consists of only one pair. When there is an intersection as in Case 2 or 3, there is at least one plateau of $\hat{f}$ or $\check{g}$ involved, which may imply that the set Γ consists of more than one pair. As we have seen in the proof of Theorem 2.17, this or these plateaus can "move"; see Fig. 2.10. Depending on whether a plateau, say of $\check{g} + s(t)$, moves up or down or stays fixed in relation to, for example, a plateau of $\hat{f}$, only one pair of the set Γ can be used in each case to obtain a solution. The correct pair is chosen by the construction procedure in §2.4. The proof of uniqueness consists of the exclusion of all other pairs of Γ. To perform these exclusions we shall use a result by Bardos, Le Roux, and Nedelec [2] concerning the two initial boundary value problems

$$\begin{aligned} &v_t + f(v)_x = 0, && x > 0,\ t > 0,\\ &v(x,0) = u_r(x), && x > 0,\\ &v^+(t) \in \tilde{N}\big(f; \alpha(t)\big), && t \geq 0, \end{aligned} \tag{2.14}$$

and

$$\begin{aligned} &w_t + g(w)_x = 0, && x < 0,\ t > 0,\\ &w(x,0) = u_l(x), && x < 0,\\ &w^-(t) \in \tilde{P}\big(g; \beta(t)\big), && t \geq 0. \end{aligned} \tag{2.15}$$

Observe the arguments f and g of $\tilde{N}$ and $\tilde{P}$. A solution of (2.14) is thus allowed to have a jump at $x = 0$ from $\alpha(t)$ to $v^+(t)$ if this discontinuity would like to move to the left, i.e., if $S\big(\alpha(t), v^+(t)\big) \equiv (f(\alpha) - f(v^+))/(\alpha - v^+) \leq 0$. Dubois and Le Floch [10] introduce the set

$$\mathcal{E}(f;\alpha) \equiv \big\{u \in \mathbb{R} : S(\alpha, k) \leq 0 \text{ for every } k \in \mathrm{ch}(u, \alpha)\big\},$$

and show the first equality in

$$\mathcal{E}(f;\alpha) = \big\{u^+(0) : u \text{ is the solution of } \mathrm{RP}(f;\alpha,\beta),\ \beta \in \mathbb{R}\big\} = \tilde{N}(f;\alpha),$$

where the last equality is implied by Lemma 2.3. They also show that $\mathcal{E}(f;\alpha)$ is equal to

$$\mathcal{E}_1(f;\alpha) \equiv \left\{u : \max_{k \in \mathrm{ch}(u,\alpha)} \operatorname{sgn}\big[(u-\alpha)\big(f(u) - f(k)\big)\big] = 0\right\}.$$

Bardos, Le Roux, and Nedelec [2] have shown that there exists a unique solution v of (2.14) that satisfies $v^+(t) \in \mathcal{E}_1\big(f; \alpha(t)\big)$. This is done by a vanishing viscosity approach. We shall use this result and the symmetrical one for (2.15) to prove that there is only one pair of functions (α, β) that satisfies Condition Γ as well as $v^+(t) = \alpha(t)$, $w^-(t) = \beta(t)$ for small $t \geq 0$. Note that $\alpha(t)$ and $\beta(t)$ are required to be continuous from the right at $t = 0$.

THEOREM 2.18. *If problem* (1.2) *is regular with g increasing, f arbitrary, and $I(\alpha,\beta,t) \neq \emptyset$, $\forall(\alpha,\beta,t) \in \mathbb{R}\times\mathbb{R}\times[0,T)$ for some $T>0$, then there exists a unique solution $u \in \Sigma$ for $t \in [0,\varepsilon)$ for some $\varepsilon \in (0,T]$.*

Remark. Note that Condition Γ is automatically fulfilled since $\check{g} \equiv$ constant for each t. A similar theorem holds for the symmetrical case when g is arbitrary and f is decreasing.

Proof. We shall only treat some cases here and we refer to [7] for the rest. The existence follows from Theorem 2.17. Let u denote the solution constructed there. Since $g'(u) > 0$, except at a discrete set of inflection points, all characteristics in $x<0$ have positive speed and define $u^-(t) = \beta(t)$ uniquely. Thus $u^-(0) = u_l(0)$ and hence $\check{g}\big(\cdot;u^-(t)\big) \equiv g\big(u^-(t)\big)$ for $t \geq 0$. It remains to prove that there is only one possibility for choosing the boundary function $\alpha(t)$, namely, the one used in the construction in Theorem 2.17. With the notation used there, define $\bar{U}(t) = \bar{U}\big(\tilde{v}^+(t),u^-(t),t\big)$ for small $t \geq 0$. Note that Lemma 2.16 gives $\bar{U}(0) = \bar{U}\big(u_r(0),u_l(0),0\big)$ and $\hat{f}\big(\bar{U}(0);\tilde{v}^+(0)\big) = \hat{f}\big(\bar{U}(0);u_r(0)\big)$. Two main cases may appear:

1. There is a unique intersection at $\bar{u}(0)$. Hence $\alpha(t)$ is uniquely determined by $\bar{u}(t) = \alpha(t)$ for small $t>0$.

2. The set $\bar{U}(0)$ is infinite, i.e., a plateau of $\hat{f}$ coincides with the constant $g\big(u^-(0)\big)+s(0)$. Depending on the graph of f, there are two main cases: A. $f\big(u_r(0)\big) \neq \hat{f}\big(\bar{U}(0);\tilde{v}^+(0)\big)$ and B. $f\big(u_r(0)\big) = \hat{f}\big(\bar{U}(0);\tilde{v}^+(0)\big)$. We shall only treat case A here and refer to [7] for all subcases that occur in case B. By symmetry it suffices to assume that $u_r(0) > \bar{u}_{\max}(0)$; see Fig. 2.11. Let α_i, $\in \{1,\dots,n\}$ be all possible u-values

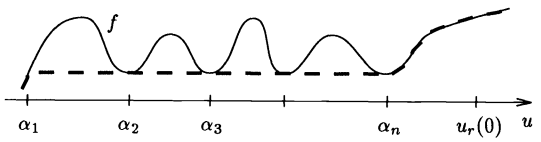


FIG. 2.11. *Case* 2A *in the proof of Theorem* 2.18. *The dashed graph is* $\hat{f}\big(\cdot;\tilde{v}^+(0)\big) = \hat{f}(\cdot;\alpha_n)$.

that satisfy $f(\alpha_i) = \hat{f}\big(\bar{U}(0);\tilde{v}^+(0)\big)$, numbered according to Fig. 2.11. Note that $\hat{f}\big(\bar{U}(0);u_r(0)\big) = \hat{f}\big(\bar{U}(0);\alpha_n\big)$. The conservation law (2.1), $f\big(\alpha(t)\big) = g\big(u^-(t)\big)+s(t)$, implies that for every admissible $\alpha(t)$, $\alpha(0) = \alpha_i$ must hold for some $i \in \{1,\dots,n\}$. Since $\tilde{v}^+(t) \equiv a = \alpha_n$ for small $t \geq 0$, which gives $f\big(\tilde{v}^+(t)\big) \equiv f(\alpha_n) = \hat{f}\big(\bar{U}(0);u_r(0)\big)$ for small $t \geq 0$, (2.10) becomes $\eta(t) = f(\alpha_n) - g\big(u^-(t)\big) - s(t)$. By the regularity assumption three cases may occur:

(i) $\eta(t) \equiv 0$ for small $t>0$. Then $\alpha(t) \equiv \alpha_i$ for small $t>0$ for some i. Independently of i, the unique solution v of (2.14) satisfies $v(x,t) = u^{\mathrm{RP}}(\frac{x}{t})$ in $\big\{(x,t) : 0<x<Kt,\ t>0\big\}$ for some $K>0$, where u^{RP} is the solution of $\mathrm{RP}\big(f;\alpha_i,u_r(0)\big)$, for $v^+(t) = u^{\mathrm{RP}}(0+) \in \tilde{N}\big(f;\alpha(t)\big)$. Thus, uniquely, $v^+(t) = u^{\mathrm{RP}}(0+) = \alpha_n = u^+(t)$ for small $t>0$.

(ii) $\eta(t) > 0$ for small $t>0$. Independently of α_i, there is a unique intersection as in 1 with $u^+(0) = \alpha_1$.

(iii) $\eta(t) < 0$ for small $t>0$. According to the construction $\alpha(0) = \alpha_n$ and $\alpha(t) > \alpha_n$ for small $t>0$. The solution is defined as in 1. Suppose instead that the boundary function $\alpha(t)$ satisfies $\alpha(0) = \alpha_i$ for some $i \in \{1,\dots,n\}$ with $\alpha(t) < \alpha_n$ and $f\big(\alpha(t)\big) = g\big(u^-(t)\big) + s(t) > f(\alpha_n)$ for small $t>0$. Then the solution v of (2.14) satisfies $v^+(t) = \alpha_n \neq \alpha(t)$. Hence the solution constructed in

Theorem 2.17 is the only possible one. □

THEOREM 2.19. *If problem* (1.2) *is regular with g decreasing, f arbitrary, and $I(\alpha,\beta,t) \neq \emptyset$, $\forall(\alpha,\beta,t) \in \mathbb{R}\times\mathbb{R}\times[0,T)$ for some $T>0$, then there exists a unique solution $u \in \Sigma$ satisfying Condition Γ for $t \in [0,\varepsilon)$ for some $\varepsilon \in (0,T]$.*

A similar theorem holds for the symmetrical case when g is arbitrary and f is increasing. The proof is found in [7].

Note that Theorem 2.18 deals with the case of intersection when $\check{g}$ is one plateau from $-\infty$ to ∞ for every t and Theorem 2.19 deals with the case when $\check{g}$ is decreasing for every t. The following theorem includes the case when a plateau with moving end point(s) is involved in the intersection. The proof is found in [7].

THEOREM 2.20. *If problem* (1.2) *is regular with $f \equiv g$ having precisely one stationary point, which is a global minimizer $u_{\min}$, then there exists a unique solution $u \in \Sigma$ satisfying Condition Γ for $t \in [0,\varepsilon)$ for some $\varepsilon > 0$.*

This theorem states that if, for example, f is convex, then there always exists a unique solution of the initial value problem for the equation $u_t + f(u)_x = s(t)\delta(x)$ (assuming regularity).

For the flux functions in the problem of continuous sedimentation (see Fig. 1.2) uniqueness in the class Σ is shown in [6].

Recall the problem of two-phase flow in porous media in §1.3. The flux functions f and g in Fig. 1.2 (right) have precisely one stationary point, which is a global minimizer. This is qualitatively the same as in Theorem 2.20 with the simplification that the source function is $s \equiv 0$. There are two qualitatively different possibilities for $\hat{f}$ and two for $\check{g}$, depending on whether $u_\pm$ lie to the left or to the right of the minimum. For a case simpler than the sedimentation problem it can be proved (see [6]) that $0 \le u_0(x) \le 1 \Rightarrow 0 \le u(x,t) \le 1$ and $I(\alpha,\beta,t) \neq \emptyset$, $\forall(\alpha,\beta,t) \in [0,1]\times[0,1]\times[0,\infty)$. Hence the procedure of construction in §2.4 can be applied if the problem is regular. Gimse and Risebro [12] have proved the existence of a global solution if the initial value u_0 has bounded total variation. They have left the question of uniqueness as an unsolved problem. The proof of Theorem 2.20 yields uniqueness in the class Σ.

2.6. The initial value boundary flux problem. Consider the *initial value boundary flux problem*

$$\begin{aligned} &u_t + f(u)_x = 0, && x>0,\ t>0,\\ &u(x,0) = u_0(x), && x>0,\\ &f\big(u^+(t)\big) = f_0(t), && t \ge 0. \end{aligned} \tag{2.16}$$

This is a variant of (1.3) when $g \equiv 0$ and $s(t) = f_0(t)$. The definitions and results above may be modified to this problem. For example, in the analogue of Theorem 2.6 no restrictions are laid on u_-. The construction of a solution can be done as in §2.4 with obvious modifications. The analysis now relies on the intersection of the graph of $\hat{f}\big(\cdot;u_+(t)\big)$ and the constant graph $f_0(t)$. Hence the proof of Theorem 2.18 also yields the following theorem. Note that Condition Γ is automatically fulfilled.

THEOREM 2.21. *If* (2.16) *is regular and $f_0(t) \in \hat{f}(\mathbb{R};\alpha)$, $\forall(\alpha,t) \in \mathbb{R}\times[0,T)$ for some $T>0$, then there exists a unique solution $u \in \Sigma$ for $t \in [0,\varepsilon)$ for some $\varepsilon \in (0,T]$.*

3. Justification of Condition Γ by Godunov's method. In this section we shall justify Condition Γ by studying a discretized version of our conservation law problem (1.2). The idea of Godunov's [13] numerical method for an equation $u_t + f(u)_x = 0$ is to use the integral form of the conservation law and the entropy

solution of the Riemann problem (1.6) to form an approximate solution by means of a discretization. The extension of this procedure to our problem (1.2), which includes the source along the t-axis, is straightforward if placing grid points *on* the t-axis. The scheme is presented in §3.1.

It is well known that Godunov's method for a scalar equation $u_t + f(u)_x = 0$ produces a sequence of approximate solutions that converges to the unique entropy satisfying solution, provided such solutions of the Riemann problem (1.6) are used in the derivation of the algorithm; see Le Roux [19]. The extension of Godunov's method to our problem does not include any extra condition along the t-axis, so it is suitable for a justification of Condition Γ. No convergence proof of the algorithm is presented here.

In §3.2 the extension of Godunov's method is used on the Riemann problem with discontinuous flux function (2.2) in two cases.

3.1. Extension of Godunov's method to problem (1.2). For δ and $\tau > 0$ let $\{(i\delta, j\tau) : i, j \in \mathbb{Z},\ j \geq 0\}$ be a grid in the half plane $\mathbb{R} \times \mathbb{R}_+$. Let

$$a = \max_{u \in M} |f'(u)| \quad \text{and} \quad b = \max_{u \in M} |g'(u)|, \tag{3.1}$$

where the interval M depends on the initial data as well as on f, g, and s; cf. Theorems 3.1 and 3.2 below. The scheme is derived by considering analytic solutions of parallel Riemann problems originating from piecewise constant initial data. These parallel solutions will not interact if choosing the ratio of the mesh size of the grid

$$\lambda \equiv \frac{\tau}{\delta} < \frac{1}{2} \min\left(\frac{1}{a}, \frac{1}{b}\right). \tag{3.2}$$

It is assumed that the ratio is constant when $\tau, \delta \searrow 0$. Using U_i^j as the approximate solution at grid point (i, j), the scheme is written as

$$U_i^{j+1} = U_i^j + \lambda\big(f(u_{i-1/2}^j) - f(u_{i+1/2}^j)\big), \quad i > 0, \tag{3.3}$$

$$U_0^{j+1} = U_0^j + \lambda\big(g(u_{-1/2}^j) - f(u_{1/2}^j) + S^j\big), \tag{3.4}$$

$$U_i^{j+1} = U_i^j + \lambda\big(g(u_{i-1/2}^j) - g(u_{i+1/2}^j)\big), \quad i < 0, \tag{3.5}$$

where

$$U_i^0 = \frac{1}{\delta} \int_{(i-1/2)\delta}^{(i+1/2)\delta} u_0(x)\, dx \qquad \text{and} \qquad S^j = \frac{1}{\tau} \int_{j\tau}^{(j+1)\tau} s(t)\, dt,$$

and from the solution of the Riemann problem (1.6) it follows that the fluxes on the straight lines in the t-direction between the grid points are given by

$$h(u_{i-1/2}^j) = \begin{cases} \min\limits_{v \in [U_{i-1}^j, U_i^j]} h(v) & \text{if } U_{i-1}^j \leq U_i^j, \\ \max\limits_{v \in [U_i^j, U_{i-1}^j]} h(v) & \text{if } U_{i-1}^j > U_i^j, \end{cases} \quad \text{where } h = \begin{cases} f, & i > 0, \\ g, & i \leq 0. \end{cases} \tag{3.6}$$

Note that $h(u_{i-1/2}^j) = \hat{h}(U_{i-1}^j, U_i^j) = \check{h}(U_i^j, U_{i-1}^j)$. Define a piecewise constant function $\tilde{U}^\tau(x,t)$ by

$$\tilde{U}^\tau(x,t) = U_i^j \quad \text{for } (x,t) \in \big[(i-1/2)\delta, (i+1/2)\delta\big) \times \big[j\tau, (j+1)\tau\big).$$

The scheme (3.3)–(3.5) is conservative in the sense that the numerical solution preserves the same amount of mass as the analytical solution. A theorem of Lax and Wendroff [16] states that if a sequence of numerical solutions obtained by a conservative method applied to the equation $u_t + f(u)_x = 0$, $x \in \mathbb{R}$, is convergent, then it converges to a weak solution. This is also true for the scheme (3.3)–(3.5) and the proof is very similar to the one of the Lax–Wendroff theorem.

3.2. Justification of Condition Γ. The justification is done in two ways. First we consider a time point when $u^{\pm}(t)$ are smooth and then we consider a time point when they are discontinuous.

Assume that the scheme (3.3)–(3.5) converges (in the weak sense) to a solution of (1.2) with $u^{\pm}$ well defined. Let t_0 be a time such that $u^{\pm}(t_0) = u_{\pm}(t_0)$, and let $j\tau \le t_0 < (j+1)\tau$ hold as $j \to \infty$, $\tau \to 0$. Hence we assume that $U^j_{\pm i} \to u^{\pm}(t_0)$, $i = 1, 2$, $U^j_0 \to u^0$, and $S^j \to s(t_0)$ as $j \to \infty$. This and the property $h(u^j_{i-1/2}) = \hat{h}(U^j_{i-1}, U^j_i) = \check{h}(U^j_i, U^j_{i-1})$ for the cells -1, 0, and 1 in the scheme (3.3)–(3.5) yield

$$\begin{aligned}
&\hat{f}(u^0; u^+) = \hat{f}(u^+; u^+),\\
&\hat{f}(u^0; u^+) = \check{g}(u^0; u^-) + s(t_0),\\
&\check{g}(u^0; u^-) = \check{g}(u^-; u^-).
\end{aligned}$$

The second equality says that $u^0 \in \bar{U}(t_0)$, and since $\hat{f}(u^+; u^+) = f(u^+)$ and $\check{g}(u^-; u^-) = g(u^-)$ hold, we conclude that

$$f(u^+) = g(u^-) + s(t_0) = \hat{f}(u^0; u^+) \quad \Longleftrightarrow \quad (u^+, u^-) \in \Gamma,$$

i.e., Condition Γ is satisfied at all time points when $u^{\pm}$ are continuous.

Now we consider the case when $u^{\pm}(t)$ are discontinuous. For a general solution of (1.2) the "most common" cases concerning the intersection of the graphs of $\hat{f}$ and $\check{g} + s(t)$ are Cases 1 and 2 after Proposition 2.11. We shall apply the scheme (3.3)–(3.5) to the Riemann problem with discontinuous flux function (2.2) (where $s \equiv 0$) in these two cases; see Theorems 3.1 and 3.2. Recall that the analytical solution, which satisfies Condition Γ, consists of two Riemann cones, one on each side of the t-axis. The boundary values on either side of the t-axis are constant for all $t \ge 0$, say $(u^+, u^-) \equiv (\alpha_0, \beta_0) = c(u_r, u_l, 0)$. Because α_0 is constant, the solution of $\mathrm{RP}(f;, \alpha_0, u_r)$ is, in $x > 0$, $t > 0$, identical to the solution of the quarter-plane problem

$$\begin{aligned}
&u_t + f(u)_x = 0, && x > 0,\ t > 0,\\
&u(x, 0) = u_r, && x > 0,\\
&u(0, t) = \alpha_0, && t \ge 0.
\end{aligned} \tag{3.7}$$

Hence the usual Godunov method produces this solution (in $x > 0$, $t > 0$) when it is applied both to $\mathrm{RP}(f; \alpha_0, u_r)$ and (3.7). Note that the constants u_r and α_0 imply that the limit of $\tilde{U}^\tau(x_0, t_0)$, when $\tau \searrow 0$, is obtained by considering the values on a fixed grid along a diagonal with speed x_0/t_0. In Theorems 3.1 and 3.2 the sequences $\{U^j_1\}$ and $\{U^j_{-1}\}$ are shown to converge to the constants that satisfy Condition Γ, i.e., $(u^+, u^-) \equiv (\alpha_0, \beta_0) = c(u_r, u_l, 0)$. Thus the sequence $\{U^j_1\}$ lies "close" to the constant sequence $\{\alpha_0\}$ and, by the reasoning above, using the latter sequence a well-defined entropy satisfying solution of (3.7) will be obtained.

THEOREM 3.1. *Assume that $f' > 0$, $g' < 0$ and that the graphs of $f \equiv \hat{f}$ and $g \equiv \check{g}$ intersect at $\bar{u}$ as in Case* 1 *(after Proposition* 2.11*). Then the sequences U_{-1}^j and U_1^j converge to $\bar{u}$ as $j \to \infty$.*

Proof. In the definition of a and b, i.e., (3.1), let M be a compact interval, which has $\bar{u}$ as the centre and contains u_l and u_r. We shall prove by induction that $U_i^j \in M$ for $i = -1, 0, 1$ so that the scheme (3.3)–(3.5) is well defined. This is true by definition for $j = 0$ because $U_0^0 = \frac{1}{2}(u_l + u_r)$ and $U_{-i}^0 = u_l$ and $U_i^0 = u_r$, $i = 1, 2, \ldots$. Assume that $U_i^j \in M$ for some $j \geq 0$, $i \in \mathbb{Z}$, then $f' > 0$, $g' < 0$, and (3.6) imply

$$\begin{aligned} &g(u_{-i-1/2}^j) = g(U_{-i}^j), \\ &f(u_{i+1/2}^j) = f(U_i^j), \qquad i = 0, 1, 2, \ldots . \end{aligned} \tag{3.8}$$

Let $h = g - f$. Then $h(\bar{u}) = 0$ and $0 > h' = g' - f' \geq -b - a$ imply

$$-(a+b) \leq \frac{h(x)}{x - \bar{u}} < 0 \quad \forall x \in M \setminus \{\bar{u}\}. \tag{3.9}$$

Since $a, b > 0$, (3.2) implies

$$\lambda < \frac{1}{2} \min\left(\frac{1}{a}, \frac{1}{b}\right) = \frac{1}{2(a+b)} \min\left(\frac{a+b}{a}, \frac{a+b}{b}\right) \leq \frac{1}{a+b},$$

which together with (3.9) gives

$$0 < 1 + \lambda \frac{h(x)}{x - \bar{u}} < 1 \quad \forall x \in M \setminus \{\bar{u}\}. \tag{3.10}$$

Using (3.8) in the iteration formula (3.4) gives

$$U_0^{j+1} - \bar{u} = \left(1 + \lambda \frac{h(U_0^j)}{U_0^j - \bar{u}}\right)(U_0^j - \bar{u}),$$

and hence (3.10) implies $U_0^{j+1} \in M$. For the 1-cell, from (3.3) we have

$$U_1^{j+1} - U_0^j = (U_1^j - U_0^j)\left(1 - \lambda \frac{f(U_1^j) - f(U_0^j)}{U_1^j - U_0^j}\right). \tag{3.11}$$

Now the fact that $f' > 0$ together with the bound (3.2) implies that the last factor in (3.11) lies strictly between $1/2$ and 1. This implies that U_1^{j+1} lies between U_1^j and U_0^j and thus $U_1^{j+1} \in M$. Repeating this procedure with the corresponding formula of (3.11) for the 2-cell, etc., we can obtain $U_i^{j+1} \in M$ for every $i = 1, 2, \ldots$. Analogously, $U_{-i}^{j+1} \in M$ holds for every $i = 1, 2, \ldots$, and the induction is finished.

The iterative formulas (3.3) and (3.4) give the discrete system

$$\begin{cases} U_0^{j+1} = U_0^j + \lambda h(U_0^j), \\ U_1^{j+1} = U_1^j + \lambda\big(f(U_0^j) - f(U_1^j)\big). \end{cases} \tag{3.12}$$

The only fixed point for this system is $(\bar{u}, \bar{u})$ and the eigenvalues of the triangular functional matrix are $1 + \lambda h'(U_0^j)$, $1 - \lambda f'(U_1^j)$. By the bound (3.2) these eigenvalues

have modulus < 1, hence $U_i^j \to \bar{u}$ as $j \to \infty$ for $i = 0, 1$. The corresponding procedure can be done with g instead of f to obtain $U_{-1}^j \to \bar{u}$ as $j \to \infty$. □

THEOREM 3.2. *Assume that $f' > 0$ and that g has precisely one stationary point, which is a global maximizer; see Fig. 2.5. Let the intersection be as in Case* 2b *(after Proposition* 2.11*) with u_l and u_2 as in Fig. 2.5. Assume that $u_l + u_r \leq 2u_2$. Then the sequences $U_{-1}^j \to u^- \equiv u_l$ and $U_1^j \to u^+ \equiv u_2$ as $j \to \infty$.*

The proof is found in [7]. The assumption $u_l + u_r \leq 2u_2$ is made to be able to apply an induction proof similar to that of Theorem 3.1. If u_r is larger we get, according to computer simulations, a transient behaviour before it is possible to apply the induction proof.

Acknowledgments. I am grateful to my supervisor Gunnar Sparr, Department of Mathematics, Lund Institute of Technology, for all of his valuable help. Thanks also to Anders Szepessy, NADA, Royal Institute of Technology, who has given significant suggestions for improvement.

REFERENCES

[1] D. P. BALLOU, *Solutions to nonlinear hyperbolic Cauchy problems without convexity conditions*, Trans. Amer. Math. Soc., 152 (1970), pp. 441–460.

[2] C. BARDOS, A. Y. LE ROUX, AND J. C. NEDELEC, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[3] T. CHANG AND L. HSIAO, *The Riemann Problem and Interaction of Waves in Gas Dynamics*, Longman Scientific and Technical, Harlow, U.K., 1989.

[4] K.-S. CHENG, *Constructing solutions of a single conservation law*, J. Differential Equations, 49 (1983), pp. 344–358.

[5] C. M. DAFERMOS, *Regularity and large time behaviour of solutions of conservation laws without convexity*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1985), pp. 201–239.

[6] S. DIEHL, *A conservation law with point source and discontinuous flux function modelling continuous sedimentation*, SIAM J. Appl. Math., to appear.

[7] ———, *On scalar conservation laws with point source and discontinuous flux function*, Tech. report ISRN LUTFD2/TFMA–7010–SE, Department of Mathematics, Lund Institute of Technology, Lund, Sweden, 1992.

[8] ———, *Scalar conservation laws with discontinuous flux function: I. The viscous profile condition*, Tech. report ISRN LUTFD2/TFMA–7005–SE, Department of Mathematics, Lund Institute of Technology, Lund, Sweden, 1994; Comm. Math. Phys., to appear.

[9] S. DIEHL AND N.-O. WALLIN, *Scalar conservation laws with discontinuous flux function: II. On the stability of the viscous profiles*, Tech. report ISRN LUTFD2/TFMA–7006–SE, Department of Mathematics, Lund Institute of Technology, Lund, Sweden, 1994; Comm. Math. Phys., to appear.

[10] F. DUBOIS AND P. LE FLOCH, *Boundary conditions for nonlinear hyperbolic systems of conservation laws*, J. Differential Equations, 71 (1988), pp. 93–122.

[11] T. GIMSE AND N. H. RISEBRO, *Riemann problems with a discontinuous flux function*, in Third International Conference on Hyperbolic Problems, Theory, Numerical Methods and Applications, Vol. I, 1990, B. Engquist and B. Gustavsson, eds., pp. 488–502.

[12] ———, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635–648.

[13] S. K. GODUNOV, *A finite difference method for the numerical computations of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb., 47 (1959), pp. 271–306. (In Russian.)

[14] S. N. KRUŽKOV, *First order quasilinear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[15] G. J. KYNCH, *A theory of sedimentation*, Trans. Faraday Soc., 48 (1952), pp. 166–176.

[16] P. D. LAX AND B. WENDROFF, *Systems of conservation laws*, Comm. Pure Appl. Math., 13 (1960), pp. 217–237.

[17] T.-P. LIU, *Nonlinear resonance for quasilinear hyperbolic equation*, J. Math. Phys., 28 (1987), pp. 2593–2602.

[18] O. A. OLEINIK, *Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation*, Uspekhi Mat. Nauk, 14 (1959), pp. 165–170; Amer. Math. Soc. Transl. Ser. 2, 33 (1964), pp. 285–290.

[19] A. Y. LE ROUX, *On the convergence of the Godunov's scheme for first order quasi linear equations*, Proc. Japan Acad. Ser. A Math. Sci., 52 (1976), pp. 488–491.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1452–1466, November 1995

004

ON THE SLOW MOTION OF VORTICES IN THE GINZBURG–LANDAU HEAT FLOW*

JACOB RUBINSTEIN† AND PETER STERNBERG‡

Abstract. We study vortex motion in the Ginzburg–Landau flow. We consider this flow in the limit of large Ginzburg–Landau parameter. It is shown that when this parameter tends to infinity, the vortex mobility tends to zero. Our proof is based on an a priori estimate on the growth of a new weighted energy and on the recent work of Bethuel, Brezis, and Helein on S^1-valued harmonic mappings in R^2.

Key words. vortices, Ginzburg–Landau heat flow

AMS subject classifications. 35Q20, 81J05

1. Introduction.

We will consider the problem of vortex motion for the Ginzburg–Landau heat flow

$$u_t = \Delta u + \varepsilon^{-2}(1 - |u|^2)u, \tag{1.1}$$

$$u(x,t) = g(x) \text{ for } x \in \partial\Omega,\ t > 0, \tag{1.2}$$

$$u(x,0) = u_0(x) \text{ for } x \in \Omega. \tag{1.3}$$

Here Ω is a two-dimensional domain, ε is a positive parameter, $u : \Omega \to \mathbf{R}^2$, $g : \partial\Omega \to \mathbf{R}^2$, and $|g| = 1$. This system appears in a canonical way when one expands a large class of second-order dissipative systems about bifurcation points [K], [BKP]. Therefore it serves as one of the fundamental models in the study of the dynamics of nonequilibrium patterns [PZM]. Equation (1.1) is also a caricature of certain models for liquid crystals [PR2].

The main objective in analyzing (1.1)–(1.3) is to study the motion of the zeroes of u. These zeroes, which are often called *vortices* or *defects*, are readily observed in many experimental setups. One would like, therefore, to understand their dynamics and equilibrium distributions.

It is easy to check that (1.1) is a gradient flow for the functional

$$\int_\Omega \frac{1}{2}|\nabla u|^2 + \frac{\varepsilon^{-2}}{4}(1 - |u|^2)^2\, dx. \tag{1.4}$$

This implies that the stable steady states of (1.1)–(1.3) are the minimizers of (1.4). A complete characterization of the minimizers in the limit $\varepsilon \to 0$ was recently announced by Bethuel, Brezis, and Helein [BBH1], [BBH2]. In particular, they have shown that the energy (1.4) is bounded from below by a term proportional to $|\ln \varepsilon|$ provided the degree of the map g is nonzero. They also found a characterization for the equilibrium location of the zeroes $(a_1, a_2, \ldots, a_d)$ of the minimizers. These zeroes

*Received by the editors November 30,1993; accepted for publication (in revised form) March 31, 1994.

†Department of Mathematics, Technion, Israel Institute of Technology, Haifa 32000, Israel. Current address: Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

‡Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

are the solution of a concrete optimization problem of minimizing an energy function $E = E(x_1, x_2, \ldots, x_d)$, where d is the degree of the boundary data g. The energy E depends only on g and the harmonic Green function of Ω.

The dynamics of the vortices in the limit $\varepsilon \to 0$ can be considered within the framework of a general program initiated by Neu [N1] and later extended by Pismen and Rubinstein [PR1]. The idea is to consider the zeroes as "particles" that interact through an external "field." The small parameter ε defines a natural length scale, and the ball of radius ε around each zero defines the particle core. In the region outside these cores the solution is dominated by a smooth field that is tempered by singularities located at the zeroes.

This program was implemented for (1.1) by several groups [BKP], [FP], [PRo], [N2], who formally used the method of matched asymptotic expansions to derive equations of motion for the zeroes and equations for their equilibrium distributions. To leading order in ε, the equations are of the form

$$m_i^{-1} \frac{d}{dt} a_i(t) = -\nabla_{a_i} E(a_1, a_2, \ldots, a_d), \quad i = 1, 2, \ldots, d. \tag{1.5}$$

The constants m_i are called the *mobilities* of the vortices.

One of the most interesting features of (1.5) that was found through these formal calculations is that the mobilities are small: $m_i \sim \frac{1}{|\ln \varepsilon|}$. Since the right-hand side of (1.5) is generically bounded (cf. [BBH2]), it follows that the vortices move *slowly,* that is, it takes a period of time which is $\mathcal{O}(|\ln \varepsilon|)$ for a vortex to move a distance 1. The main result in this paper is a theorem stating that under certain conditions, the mobilities are indeed $\mathcal{O}(\frac{1}{|\ln \varepsilon|})$. The theorem is formulated and proved in the next two sections. The proof relies on a weighted energy estimate derived from the parabolic Pohozaev identity obtained in [BCPS]. A key observation is that the weighted energy (see (2.5) below) remains bounded uniformly in ε for data possessing a vortex at the origin. The proof makes crucial use of the lower bound for the energy of a minimizer of (1.4) found in [BBH2].

We have quoted the results of Bethuel, Brezis, and Helein, who showed that the number of zeroes in the steady state is precisely the degree of the boundary data. This fact has also been demonstrated, in some set-ups, for arbitrary (finite) values of ε [BCP], [BCPS]. By degree-theoretic considerations, this is the minimal number of zeroes that a smooth function must have to be compatible with g. This does not preclude, however, the possibility that the number of zeroes increases in the course of the evolution (1.1)–(1.3). In §4 we demonstrate, through an explicit construction, that an arbitrary number of vortices might emerge spontaneously. Clearly, these vortices appear in pairs with degree ± 1.

2. Main result. We take $\Omega \subset \mathbf{R}^2$ to be a bounded convex domain with smooth boundary. Then, consider smooth boundary data $g : \partial\Omega \to \mathbf{R}^2$ satisfying the conditions

$$|g| = 1 \tag{2.1}$$

and

$$\deg(g, \partial\Omega) = 1. \tag{2.2}$$

Here $\deg(g, \partial\Omega)$ denotes the Brouwer degree (i.e., the winding number of g considered as a map from $\partial\Omega$ into S^1). We take the initial data u_0 $(= u_0^\varepsilon)$ to be a smooth mapping

from Ω to $\mathbf{R}^2$ satisfying the following conditions:

$$|u_0| \leq 1, \tag{2.3}$$

$$E_\varepsilon(u_0) \equiv \int_\Omega 1/2\,|\nabla u_0|^2 + \varepsilon^{-2} V(u_0)\,dx \leq \pi\,|\ln \varepsilon| + C_0, \tag{2.4}$$

$$E_\varepsilon^w(u_0) \equiv \int_\Omega |x|^2\,[1/2\,|\nabla u_0|^2 + \varepsilon^{-2} V(u_0)]\,dx \leq C_1, \tag{2.5}$$

$$\int_\Omega \left|\Delta u_0 + \varepsilon^{-2}(1 - |u_0|^2)u_0\right|^2 dx \leq C_2 \frac{|\ln \varepsilon|}{\varepsilon^2}, \tag{2.6}$$

where $V(u) = \frac{1}{4}(|u|^2 - 1)^2$. Furthermore, we assume u_0 is compatible with g. For simplicity, we will assume the data is first-order compatible to appeal to standard existence, uniqueness, and regularity theory, but this is by no means necessary. Finally, we take the origin to be the only zero of u_0 in Ω with, necessarily,

$$\deg(u_0, \partial B_r(0)) = 1 \tag{2.7}$$

for all $r < \operatorname{dist}(0, \partial\Omega)$. Here and throughout the paper, $B_r(x)$ denotes the ball of radius r centered at x. The constants C_0, C_1, and C_2 above are taken to be independent of ε.

In §3 we will explicitly construct initial data satisfying these conditions.

Remark. We shall only need the fact that Ω is star shaped with respect to the original vortex location—placed for convenience at the origin. However, since we wish to allow the vortex to be originally located anywhere in Ω, this leads to the assumption of convexity.

Now let u_ε denote the solution to problem (1.1)–(1.3), where g and u_0 satisfy conditions (2.1)–(2.7). The existence of a unique classical solution to (1.1)–(1.3) follows from Theorem 7.1 of [LSU, Chap. VII]. Furthermore, applying the maximum principle to the differential equation satisfied by $|u_\varepsilon|^2$, one finds that (2.1) and (2.3) imply

$$|u_\varepsilon| < 1 \text{ for } x \in \Omega,\ t > 0. \tag{2.8}$$

We now state our main result.

THEOREM 2.1. *Assume that at each time $t > 0$ there exists exactly one zero of u_ε, denoted by $q_\varepsilon(t)$, with $\deg(u_\varepsilon, \partial B_r(q_\varepsilon(t))) = 1$ for all positive $r < \operatorname{dist}(q_\varepsilon(t), \partial\Omega)$. Let R and λ be any postive numbers and denote by T_ε the infimum of the set $\{t : |q_\varepsilon(t)| = R,\ \operatorname{dist}(q_\varepsilon(t), \partial\Omega) \geq \lambda\}$, assuming this set to be nonempty. Then*

$$\liminf_{\varepsilon \to 0} \frac{T_\varepsilon}{|\ln \varepsilon|} > 0. \tag{2.9}$$

To prove (2.9) we will need the following two results.

LEMMA 2.2 (weighted energy estimate). *There exists a constant C_3 independent of ε such that*

$$E_\varepsilon^w(u_\varepsilon)(T) \leq E_\varepsilon^w(u_0) + C_3 T$$

for all $T > 0$.

LEMMA 2.3. *For any $\alpha \in (0,1)$ we have the spatial Hölder estimate*

$$\sup_{x,y \in \Omega} \frac{|u_\varepsilon(x,t) - u_\varepsilon(y,t)|}{|x-y|^\alpha} \leq C_4 \frac{\sqrt{|\ln \varepsilon|}}{\varepsilon} \tag{2.10}$$

for some positive constant C_4 independent of ε and t.

We will begin with the proofs of the lemmas and then prove our theorem.

Proof of Lemma 2.2. For simplicity of presentation, in this proof we will suppress the dependence of the solution on ε and write u for u_ε. Since Ω is star shaped with respect to the origin, there exists a number $\alpha_0 > 0$ such that

$$x \cdot \nu \geq \alpha_0 \quad \text{for all } x \in \partial\Omega, \tag{2.11}$$

where ν denotes the outer unit normal to $\partial\Omega$. Proceeding as in the proof of Lemma 4.1 in [BCPS], we first take the inner product of (1.1) with $\nabla u \cdot x$ and integrate over Ω. After repeated integration by parts and using $V(u) = 0$ for $x \in \partial\Omega$, one obtains

$$\int_\Omega (u_t \cdot (\nabla u \cdot x))\, dx + \int_{\partial\Omega} 1/2(x\cdot\nu)\,|\nabla u|^2 - (\nabla u\cdot\nu)(\nabla u\cdot x)\, ds = 2\varepsilon^{-2}\int_\Omega V(u)\, dx. \tag{2.12}$$

Now take the inner product of (1.1) with the quantity $1/2\,|x|^2\, u_t$ and integrate over Ω. After integrating by parts and using (1.2) to conclude that $u_t = 0$ on $\partial\Omega$, we find

$$\frac{d}{dt}\int_\Omega |x|^2\,[1/2\,|\nabla u|^2 + \varepsilon^{-2}V(u)]\, dx = -\int_\Omega 1/2\,|x|^2\,|u_t|^2 + (\nabla u \cdot x)\cdot u_t\, dx. \tag{2.13}$$

Substitution of (2.12) into (2.13) yields

(2.14)

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega |x|^2\,[1/2\,|\nabla u|^2 + \varepsilon^{-2}V(u)]\, dx \\
&\quad = -\int_\Omega [1/2\,|x|^2\,|u_t|^2 + 2\varepsilon^{-2}V(u)]\, dx + \int_{\partial\Omega} 1/2(x\cdot\nu)\,|\nabla u|^2 - (\nabla u\cdot\nu)(\nabla u\cdot x)\, ds \\
&\quad \leq -\int_{\partial\Omega} 1/2(x\cdot\nu)(\nabla u\cdot\nu)^2 + (x\cdot\tau)(\nabla u\cdot\nu)(\nabla u\cdot\tau)\, ds + 1/2\int_{\partial\Omega}(x\cdot\nu)\,|\partial_\tau g|^2\, ds,
\end{aligned}$$

where τ denotes the unit tangent to $\partial\Omega$ and $\partial_\tau g$ denotes the derivative of the boundary data along $\partial\Omega$. We note that (2.14) is a parabolic version of the well-known Pohozaev identity from elliptic partial differential equations. Other such identities were discovered earlier by Giga and Kohn; see [GK].

Utilizing (2.11), we conclude that

$$\frac{d}{dt}\int_\Omega |x|^2\,[1/2\,|\nabla u|^2 + \varepsilon^{-2}V(u)]\, dx \leq C_3$$

for some constant C_3 depending on α_0, Ω, and g. The result follows. □

Proof of Lemma 2.3. First note that by direct calculation we find that the energy E_ε is nonincreasing along the flow

$$\frac{d}{dt}\int_\Omega 1/2\,|\nabla u|^2 + \varepsilon^{-2}V(u)\, dx = -\int_\Omega |u_t|^2\, dx \leq 0. \tag{2.15}$$

Hence, by (2.4),

$$E_\varepsilon(u)(t) \leq E_\varepsilon(u_0) \leq \pi\,|\ln\varepsilon| + C_0. \tag{2.16}$$

We use this and (2.8) to conclude that

$$\begin{aligned}\|\varepsilon^{-2}(|u|^2-1)u\|_{L^2(\Omega)} &= \frac{1}{\varepsilon}\left(\int_\Omega \varepsilon^{-2}(|u|^2-1)^2\,|u|^2\,dx\right)^{1/2}\\ &\le \frac{1}{\varepsilon}\left(\int_\Omega \varepsilon^{-2}V(u)\,dx\right)^{1/2}\\ &\le \frac{\sqrt{\pi\,|\ln\varepsilon|+C_0}}{\varepsilon}\\ &\le C\frac{\sqrt{|\ln\varepsilon|}}{\varepsilon}\end{aligned} \tag{2.17}$$

for some constant C independent of ε and t.

Now differentiate equation (1.1) with respect to time, take the inner product with the quantity $2u_t$, and integrate over Ω to find

$$\int_\Omega u_t\cdot u_{tt}\,dx = \int_\Omega u_t\cdot\Delta(u_t)+\varepsilon^{-2}(1-|u|^2)\,|u_t|^2-2\varepsilon^{-2}(u\cdot u_t)^2\,dx.$$

Thus

$$\frac{d}{dt}\int_\Omega \frac{1}{2}\,|u_t|^2\,dx = -\int_\Omega\left(|\nabla u_t|^2+\varepsilon^{-2}\,|u|^2\,|u_t|^2+2\varepsilon^{-2}(u\cdot u_t)^2\right)dx+\varepsilon^{-2}\int_\Omega |u_t|^2\,dx. \tag{2.18}$$

Multiplying (2.15) by ε^{-2} and adding it to (2.18) we find

$$\begin{aligned}&\frac{d}{dt}\int_\Omega \frac{1}{2}\,|u_t|^2+\varepsilon^{-2}\left(1/2\,|\nabla u|^2+\varepsilon^{-2}V(u)\right)dx\\ &\qquad = -\int_\Omega\left(|\nabla u_t|^2+\varepsilon^{-2}\,|u|^2\,|u_t|^2+2\varepsilon^{-2}(u\cdot u_t)^2\right)dx\le 0.\end{aligned}$$

Invoking hypotheses (2.4) and (2.6), we conclude that

$$\int_\Omega \frac{1}{2}\,|u_t(x,t)|^2\,dx+\varepsilon^{-2}E_\varepsilon(u)(t)\le\int_\Omega\frac{1}{2}\left|\Delta u_0+\varepsilon^{-2}(1-|u_0|^2)u_0\right|^2dx+\varepsilon^{-2}E_\varepsilon(u_0) \tag{2.19}$$

for some constant C independent of ε and t. We now appeal to assumptions (2.4) and (2.6) to obtain

$$\|u_t\|_{L^2(\Omega)}\le C\frac{\sqrt{|\ln\varepsilon|}}{\varepsilon} \tag{2.20}$$

for some constant C. In view of (1.1), estimates (2.17) and (2.20) imply that for all $t\ge 0$,

$$\|\Delta u\|_{L^2(\Omega)}\le C\frac{\sqrt{|\ln\varepsilon|}}{\varepsilon}.$$

Hence, by Theorem 8.12 of [GT], we have

$$\|u\|_{W^{2,2}(\Omega)}\le C\frac{\sqrt{|\ln\varepsilon|}}{\varepsilon},$$

and the Sobolev imbedding theorem implies (2.10). We note that this result is not sharp; indeed, based on gradient estimates for the elliptic case, one would expect that the L^∞ norm of the gradient is bounded by C/ε, but (2.10) suffices for our purposes. □

Proof of Theorem 2.1.

Step 1. Fix any $R > 0$ and $\lambda > 0$ and recall that we define the time T_ε by

$$T_\varepsilon = \inf\{t : |q_\varepsilon(t)| = R,\ \operatorname{dist}(q_\varepsilon(t), \partial\Omega) \geq \lambda\}, \tag{2.21}$$

where we assume this set is nonempty. For convenience set

$$a \equiv \frac{1}{4}\min(R, \lambda).$$

Now define the set

$$S_\varepsilon = \left\{ r \in (a, 2a) : |u_\varepsilon(x, T_\varepsilon)| \geq \frac{1}{4} \text{ for all } x \text{ such that } |x - q_\varepsilon(T_\varepsilon)| = r \right\}. \tag{2.22}$$

We first will argue that there exists a number $\varepsilon_0 > 0$ such that

$$H^1(S_\varepsilon) \geq \frac{a}{2} \tag{2.23}$$

for all $\varepsilon < \varepsilon_0$, where $H^1(S_\varepsilon)$ denotes the (one-dimensional) measure of the set S_ε. We proceed by contradiction and suppose there exists a sequence $\{\varepsilon_j\} \to 0$ such that

$$H^1(S^c_{\varepsilon_j}) > \frac{a}{2}, \tag{2.24}$$

where $S^c_{\varepsilon_j}$ denotes the complement of S_{ε_j} in the interval $(a, 2a)$, i.e., $S^c_{\varepsilon_j} = (a, 2a) - S_{\varepsilon_j}$. By definition (2.22), we note that for all $r \in S^c_{\varepsilon_j}$ there exists a point x_r $(= x_r(\varepsilon_j))$ satisfying

$$\left|x_r - q_{\varepsilon_j}(T_{\varepsilon_j})\right| = r, \quad \left|u_{\varepsilon_j}(x_r, T_{\varepsilon_j})\right| < \frac{1}{4}. \tag{2.25}$$

For simplicity of notation, in the remainder of Step 1 we will suppress the subsequential index j and simply write ε for ε_j. Invoking Lemma 2.3 with, say, $\alpha = 3/4$, we find that for all $r \in S^c_\varepsilon$ we have

$$\begin{aligned} |u_\varepsilon(x, T_\varepsilon)| &\leq |u_\varepsilon(x_r, T_\varepsilon)| + |u_\varepsilon(x, T_\varepsilon) - u_\varepsilon(x_r, T_\varepsilon)| \\ &\leq \frac{1}{4} + C_4 R_\varepsilon^{3/4} \frac{\sqrt{|\ln \varepsilon|}}{\varepsilon} \leq \frac{1}{2} \end{aligned} \tag{2.26}$$

for all x satisfying $|x - x_r| \leq R_\varepsilon$, where

$$R_\varepsilon \equiv \left(\frac{\varepsilon}{4C_4\sqrt{|\ln \varepsilon|}} \right)^{4/3}. \tag{2.27}$$

Hence, for all $r \in S^c_\varepsilon$ and $x \in B_{R_\varepsilon}(x_r)$ we have

$$\begin{aligned} V(u_\varepsilon(x, T_\varepsilon)) &= \frac{1}{4}(1 - |u|)^2(1 + |u|)^2 \\ &\geq \frac{1}{4}(1 - |u|)^2 \geq \frac{1}{16}, \end{aligned} \tag{2.28}$$

from which it follows that

$$\varepsilon^{-2}\int_\Omega V(u_\varepsilon(x,T_\varepsilon))\,dx \geq \varepsilon^{-2}\int_{\cup_{r\in S_\varepsilon^c} B_{R_\varepsilon}(x_r)} V(u_\varepsilon(x,T_\varepsilon))\,dx$$
$$\geq \frac{1}{16}\varepsilon^{-2}H^2(\{\cup_{r\in S_\varepsilon^c} B_{R_\varepsilon}(x_r)\}),$$

where H^2 denotes the two-dimensional measure. In light of (2.24), we find

$$H^2(\{\cup_{r\in S_\varepsilon^c} B_{R_\varepsilon}(x_r)\}) \geq CR_\varepsilon$$

for some constant C independent of ε and t. We then use (2.16) to reach the desired contradiction

$$\pi\,|\ln\varepsilon| + C_0 \geq E(u_0) \geq E(u_\varepsilon)(T_\varepsilon)$$
$$\geq \varepsilon^{-2}\int_\Omega V(u_\varepsilon(x,T_\varepsilon))\,dx \geq \frac{1}{16}\varepsilon^{-2}CR_\varepsilon = \frac{C}{16}\varepsilon^{-2}\left(\frac{\varepsilon}{4C_4\sqrt{|\ln\varepsilon|}}\right)^{4/3}.$$

Step 2. Next we will establish the estimate

$$\int_{B_{2a}(q_\varepsilon(T_\varepsilon))} \frac{1}{2}\,|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx \geq C_5\,|\ln\varepsilon| - C_6 \tag{2.29}$$

for positive constants C_5 and C_6 independent of ε. To this end, we first claim that by (2.23), for every positive $\varepsilon < \varepsilon_0$ we can find a number $r_\varepsilon \in S_\varepsilon$ such that

$$\int_{\partial B_{r_\varepsilon}} \frac{1}{2}\,|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,ds$$
$$\leq \frac{3}{a}\int_{B_{2a}} \frac{1}{2}\,|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx. \tag{2.30}$$

In (2.30) and throughout Step 2 all balls will be centered at the vortex location $q_\varepsilon(T_\varepsilon)$ and we shall simply write B_r for the ball centered at $q_\varepsilon(T_\varepsilon)$ of radius r. If (2.30) were false for all $r \in S_\varepsilon$, (2.23) would immediately lead to the contradiction

$$\int_{B_{2a}} \frac{1}{2}\,|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx$$
$$\geq \int_{r\in S_\varepsilon}\int_{\partial B_r} \frac{1}{2}\,|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,ds\,dr$$
$$> \frac{3}{a}H^1(S_\varepsilon)\int_{B_{2a}} \frac{1}{2}\,|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx$$
$$\geq \frac{3}{2}\int_{B_{2a}} \frac{1}{2}\,|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx.$$

We will establish estimate (2.29) by comparing the Dirichlet integral of $u_\varepsilon(\cdot,T_\varepsilon)$ in B_{3a} to the Dirichlet integral of a constructed function which agrees with u_ε inside B_{r_ε} but which equals $(x - q_\varepsilon(T_\varepsilon))/|x - q_\varepsilon(T_\varepsilon)|$ outside $B_{r_\varepsilon+a}$.

To facilitate this construction, we now introduce a local polar coordinate system (r,θ) centered at the vortex $q_\varepsilon(T_\varepsilon)$. Since $\deg(u_\varepsilon(\cdot,T_\varepsilon),\partial B_r)=1$ for all $r\in(0,4a)$, we can, in particular, write

$$u_\varepsilon(r_\varepsilon,\theta,T_\varepsilon)=\zeta_\varepsilon(\theta)e^{i(\theta+\phi_\varepsilon(\theta))},$$

where ϕ_ε is a smooth function such that $\phi_\varepsilon(0)=\phi_\varepsilon(2\pi)=0$, and by the definition (2.22) of S_ε, ζ_ε is a smooth function satisfying

$$\zeta_\varepsilon(0)=\zeta_\varepsilon(2\pi),\quad \frac{1}{4}\le\zeta_\varepsilon(\theta)\le 1 \text{ for all } \theta\in[0,2\pi]. \tag{2.31}$$

Now define the functions A_ε and β_ε by the formulas

$$A_\varepsilon(r,\theta)=\left(\frac{1-\zeta_\varepsilon(\theta)}{a}\right)(r-r_\varepsilon)+\zeta_\varepsilon(\theta) \text{ for } r\in(r_\varepsilon,r_\varepsilon+a),\ \theta\in[0,2\pi]$$

and

$$\beta_\varepsilon(r,\theta)=-\frac{\phi_\varepsilon(\theta)}{a}(r-r_\varepsilon)+\phi_\varepsilon(\theta) \text{ for } r\in(r_\varepsilon,r_\varepsilon+a),\ \theta\in[0,2\pi],$$

and let $w_\varepsilon:\Omega\to\mathbf{R}^2$ be defined by

$$w_\varepsilon(r,\theta)=\begin{cases} u_\varepsilon(r,\theta,T_\varepsilon) & \text{for } r\le r_\varepsilon,\\ A_\varepsilon(r,\theta)e^{i(\theta+\beta_\varepsilon(r,\theta))} & \text{for } r_\varepsilon<r<r_\varepsilon+a,\\ e^{i\theta} & \text{for } r\ge r_\varepsilon+a.\end{cases} \tag{2.32}$$

Note that w_ε is a continuous $H^1(\Omega)$ function.

We now appeal to the lower bound on the energy of a minimizer (cf. Theorem 5 of [BBH2]), which says

$$\inf_v E_\varepsilon(v)\ge\pi|\ln\varepsilon|-C \tag{2.33}$$

for some constant C independent of ε (but depending on the domain and the degree 1 boundary data), where the minimization is taken over H^1 functions with given boundary data. We therefore conclude that

$$\begin{aligned}\pi|\ln\varepsilon|-C&\le\int_{B_{3a}}\frac{1}{2}|\nabla w_\varepsilon|^2+\varepsilon^{-2}V(w_\varepsilon)\,dx\\ &=\int_{B_{3a}-B_{r_\varepsilon+a}}\cdot+\int_{B_{r_\varepsilon+a}-B_{r_\varepsilon}}\cdot+\int_{B_{r_\varepsilon}}\cdot\end{aligned} \tag{2.34}$$

The first of these three integrals is independent of ε, since w_ε is independent of ε there and $V(w_\varepsilon)=V(e^{i\theta})=0$. For the last of the three integrals $w_\varepsilon=u_\varepsilon$, so it is certainly bounded by

$$\int_{B_{2a}}\frac{1}{2}|\nabla u_\varepsilon(x,T_\varepsilon)|^2+\varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx.$$

Also, in the second integral we have

$$\int_{B_{r_\varepsilon+a}-B_{r_\varepsilon}}\varepsilon^{-2}V(w_\varepsilon)\,dx\le\int_0^{2\pi}\int_{r_\varepsilon}^{r_\varepsilon+a}\varepsilon^{-2}V(u_\varepsilon(r_\varepsilon,\theta,T_\varepsilon))r\,dr\,d\theta,$$

since

$$|u_\varepsilon(r_\varepsilon,\theta,T_\varepsilon)| = \zeta_\varepsilon(\theta) \le A_\varepsilon(r,\theta) = |w_\varepsilon(r,\theta)|$$

for $r_\varepsilon \le r \le r_\varepsilon + a$. It then follows from (2.30) that

$$\int_{B_{r_\varepsilon+a}-B_{r_\varepsilon}} \varepsilon^{-2}V(w_\varepsilon)\,dx \le C\int_{B_{2a}} \frac{1}{2}|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx$$

for some constant C independent of ε. In view of (2.34), (2.29) will then be established if

(2.35)
$$\int_{B_{r_\varepsilon+a}-B_{r_\varepsilon}} |\nabla w_\varepsilon|^2\,dx \le \text{const.}\int_{B_{2a}} |\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx + \text{const.}$$

Using the definition of w_ε in (2.32) we find

$$\begin{aligned}
\int_{B_{r_\varepsilon+a}-B_{r_\varepsilon}} |\nabla w_\varepsilon|^2\,dx &= \int_0^{2\pi}\int_{r_\varepsilon}^{r_\varepsilon+a}\left(\left|\frac{\partial w_\varepsilon}{\partial r}\right|^2 + \frac{1}{r^2}\left|\frac{\partial w_\varepsilon}{\partial\theta}\right|^2\right) r\,dr\,d\theta \\
&= \int_0^{2\pi}\int_{r_\varepsilon}^{r_\varepsilon+a}\left(\left|\frac{\partial A_\varepsilon}{\partial r}\right|^2 + \left|A_\varepsilon\frac{\partial\beta_\varepsilon}{\partial r}\right|^2\right) r\,dr\,d\theta \\
&\quad + \int_0^{2\pi}\int_{r_\varepsilon}^{r_\varepsilon+a}\left(\left|\frac{\partial A_\varepsilon}{\partial\theta}\right|^2 + \left|A_\varepsilon\left(1+\frac{\partial\beta_\varepsilon}{\partial\theta}\right)\right|^2\right)\frac{1}{r}\,dr\,d\theta \\
&\equiv I_1 + I_2.
\end{aligned} \tag{2.36}$$

Writing estimate (2.30) in terms of the L^2-norms of $d\zeta_\varepsilon/d\theta$ and $d\phi_\varepsilon/d\theta$, we get bounds on these norms in terms of the energy of u_ε in B_{2a}, as well as an L^2-bound on ϕ_ε itself, via the Poincare inequality. In light of the uniform boundedness of $|\zeta_\varepsilon|$ and $|A_\varepsilon|$, integral I_1 is then bounded in the sense of (2.35). Similarly, one bounds I_2 to obtain (2.35) and (2.29) follows.

Step 3. Invoking (2.29), we find that

$$\begin{aligned}
E_\varepsilon^w(u_\varepsilon)(T_\varepsilon) &= \int_\Omega |x|^2\left[\frac{1}{2}|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\right]dx \\
&\ge \int_{B_{2a}(q(T_\varepsilon))} |x|^2\left[\frac{1}{2}|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\right]dx \\
&\ge 4a^2\int_{B_{2a}(q(T_\varepsilon))}\frac{1}{2}|\nabla u_\varepsilon(x,T_\varepsilon)|^2 + \varepsilon^{-2}V(u_\varepsilon(x,T_\varepsilon))\,dx \\
&\ge 4a^2(C_5|\ln\varepsilon| - C_6).
\end{aligned} \tag{2.37}$$

Then, appealing to (2.5) and Lemma 2.2, we obtain

$$\begin{aligned}
4a^2(C_5|\ln\varepsilon| - C_6) &\le E_\varepsilon^w(u_\varepsilon)(T_\varepsilon) \\
&\le E_\varepsilon^w(u_0) + C_3T_\varepsilon \le C_1 + C_3T_\varepsilon
\end{aligned}$$

and Theorem 2.1 follows. □

3. Construction of initial data. We will show in this section, through an explicit construction, that there exists initial data $u_0 = u_0^\varepsilon$ satisfying the bounds

(2.3)–(2.6), which has exactly one zero at the origin satisfying (2.7), and which is compatible with the given data g. Thus it is valid data to use in applying Theorem 2.1. While the construction is quite special, suggesting that application of Theorem 2.1 may be limited, we wish to point out that these assumptions really amount to ignoring transient behavior in the solution u_ε to (1.1)–(1.3) for arbitrary initial data. That is, one can argue formally that the solution with arbitrary data will rapidly develop into a function meeting the criteria for data given in Theorem 2.1.

We will define our candidate for u_0 in terms of a polar coordinate system (r, θ) centered at the origin. In view of (2.2), we may write

$$g = g(\theta) = e^{i(\theta + \chi(\theta))}, \tag{3.1}$$

where χ is a smooth function such that $\chi(0) = \chi(2\pi)$. The construction will make use of the separation of variables equilibrium solution to (1.1) (with $\varepsilon = 1$) given by $\rho(s)e^{i\theta}$, where ρ satisfies the ordinary differential equation

$$-\rho'' - \frac{1}{s}\rho' + \frac{1}{s^2}\rho = \rho(1 - \rho^2) \text{ for } 0 < s < \infty \tag{3.2}$$

($' = \frac{d}{ds}$), subject to the boundary conditions

$$\rho(0) = 0, \quad \rho(\infty) = 1. \tag{3.3}$$

It is known that (3.2), (3.3) admits a unique solution which has the asymptotic behavior

$$\rho(s) \sim 1 - \frac{1}{2s^2}, \quad \rho'(s) \sim \frac{1}{s^3} \quad \text{as } s \to \infty. \tag{3.4}$$

Denoting

$$d \equiv \frac{1}{3} \inf_{x \in \partial\Omega} |x|,$$

we let U_ε be the solution to the variational problem

$$\inf_u \int_{\Omega - B_d(0)} \frac{1}{2} |\nabla u|^2 + \frac{\varepsilon^{-2}}{4}(1 - |u|^2)^2 \, dx, \tag{3.5}$$

where the minimization is taken over functions in $H^1(\Omega - B_d(0))$ satisfying the Dirichlet conditions

$$u = g \text{ on } \partial\Omega, \quad u = \rho\left(\frac{d}{\varepsilon}\right) e^{i\theta} \text{ on } \partial B_d(0). \tag{3.6}$$

The existence of a (smooth) solution to (3.5), (3.6) follows easily by the direct method. We then have the following proposition.

PROPOSITION 3.1. *The function $u_0 = u_0^\varepsilon$ defined on Ω by*

$$u_0 = \begin{cases} \rho(\frac{|x|}{\varepsilon}) & \text{for } |x| \le d, \\ U_\varepsilon(x) & \text{for } |x| > d \end{cases}$$

satisfies (2.3)–(2.7). *Furthermore, it is first-order compatible.*

Remark. While u_0 so defined is not smooth on the circle $r = d$, it is clear that it can be smoothed out in a neighborhood of this circle while still satisfying (2.3)–(2.7), so we shall ignore this issue.

Proof. Property (2.3) follows readily from the maximum principle applied to the quantity $|U_\varepsilon|^2$, while property (2.7) and the compatibility are an immediate consequence of the definition of u_0.

Verification of (2.4). We write

$$\begin{aligned} E_\varepsilon(u_0) &= \int_\Omega \frac{1}{2}|\nabla u_0|^2 + \varepsilon^{-2}V(u_0)\,dx \\ &= \int_{\{r<d\}} \cdot + \int_{\{r>d\}} \end{aligned} \tag{3.7}$$

Now we compute

$$\begin{aligned} \int_{\{r<d\}} \cdot &= \int_0^{2\pi}\int_0^d \left[\frac{1}{2}\left|\frac{\partial u_0}{\partial r}\right|^2 + \frac{1}{2r^2}\left|\frac{\partial u_0}{\partial \theta}\right|^2 + \varepsilon^{-2}(\rho^2-1)^2\right] r\,dr\,d\theta \\ &= 2\pi\int_0^{d/\varepsilon}\left[\frac{s}{2}|\rho'(s)|^2 + \frac{1}{2s}\rho(s)^2 + s(\rho(s)^2-1)^2\right] ds. \end{aligned} \tag{3.8}$$

Using (3.3) and (3.4) one easily checks that

$$2\pi\int_0^{d/\varepsilon} \frac{s}{2}|\rho'(s)|^2 + s(\rho(s)^2-1)^2\,ds + 2\pi\int_0^1 \frac{1}{2s}\rho(s)^2\,ds \le C \tag{3.9}$$

for some constant C independent of ε, while

$$\pi\int_1^{d/\varepsilon} \frac{1}{s}\rho(s)^2\,ds \le C + \pi|\ln\varepsilon|\,. \tag{3.10}$$

To bound $\int_{\{r>d\}} \cdot$, we introduce the competitor $\Psi_\varepsilon(r)e^{i(\theta+\Phi_\varepsilon(r,\theta))}$ in the variational problem (3.5), (3.6), where

$$\Psi_\varepsilon(r) = \begin{cases} \left[\frac{1-\rho(\frac{d}{\varepsilon})}{d}\right](r-2d)+1 & \text{for } d \le r \le 2d, \\ 1 & \text{for } r > 2d \end{cases} \tag{3.11}$$

and

$$\Phi_\varepsilon(r,\theta) = \begin{cases} \frac{\chi(\theta)}{d}(r-d) & \text{for } d \le r \le 2d, \\ \chi(\theta) & \text{for } r > 2d. \end{cases} \tag{3.12}$$

We claim that the energy of this competitor in the set $\Omega - B_d(0)$ is bounded by a constant independent of ε, from which it will follow that the same is true of U_ε, thus establishing (2.4). To this end, note that since $\Psi_\varepsilon(r)e^{i(\theta+\Phi_\varepsilon(r,\theta))}$ is a unit vector independent of ε in the set $\{r > 2d\}$, this claim will be proven if one can bound the energy of the competitor in the set $\{d < r < 2d\}$. A direct calculation using (3.4), (3.11), and (3.12) yields such a bound; we omit the details.

Verification of (2.5). We write

$$\begin{aligned} E_\varepsilon^w(u_0) &= \int_\Omega |x|^2\left[\frac{1}{2}|\nabla u_0|^2 + \varepsilon^{-2}V(u_0)\right] dx \\ &= \int_{\{r<d\}} \cdot + \int_{\{r>d\}} \cdot \end{aligned} \tag{3.13}$$

In the verification of (2.4) we found that

$$\int_{\{r>d\}} \frac{1}{2}|\nabla u_0|^2 + \varepsilon^{-2}V(u_0)\,dx < C,$$

so

$$\int_{\{r>d\}} |x|^2 \left[\frac{1}{2}|\nabla u_0|^2 + \varepsilon^{-2}V(u_0)\right] dx$$

is bounded independent of ε as well. Then we compute

$$\begin{aligned}
&\int_{\{r<d\}} |x|^2 \left[\frac{1}{2}|\nabla u_0|^2 + \varepsilon^{-2}V(u_0)\right] dx \\
&\quad = \int_0^{2\pi}\int_0^d r^2 \left[\frac{1}{2}\left|\frac{\partial u_0}{\partial r}\right|^2 + \frac{1}{2r^2}\left|\frac{\partial u_0}{\partial \theta}\right|^2 + \varepsilon^{-2}(\rho^2-1)^2\right] r\,dr\,d\theta \qquad (3.14)\\
&\quad = 2\pi\int_0^{d/\varepsilon} \left[\frac{s}{2}|\rho'(s)|^2 + \frac{1}{2s}\rho(s)^2 + s(\rho(s)^2-1)^2\right]\varepsilon^2 s^2\,ds.
\end{aligned}$$

In light of (3.9), the only relevant term to check here is

$$\pi\varepsilon^2\int_0^{d/\varepsilon} s\rho(s)^2\,ds \le \pi\varepsilon^2\int_0^{d/\varepsilon} s\,ds \le C,$$

thus completing the verification of (2.5).

Verification of (2.6). We write

$$\int_\Omega \left|\Delta u_0 - \varepsilon^{-2}(1-|u_0|^2)u_0\right|^2 dx = \int_{\{r<d\}} \cdot + \int_{\{r>d\}} \cdot$$

Since u_0 is an exact equilibrium solution of (1.1) in the regions $\{r<d\}$ and $\{r>d\}$, we find that both integrals are zero. Of course, after smoothing our construction near $\{r=d\}$, there will be some positive contribution to the integral, but this process can be easily carried out while only creating a contribution that is $\mathcal{O}(1)$. We note that while (2.6) will suffice in proving Theorem 2.1, our construction yields a much better bound. □

4. Creation of vortices. In this section we will present an example which shows that it is possible for zeroes (vortices) of degree ± 1 to emerge after finite time in the solution u_ε to (1.1)–(1.3). The reader will note that in Theorem 2.1 we assume that only one zero (of degree 1) exists for all time. It is because of the example presented below that we must make such an assumption. While we firmly believe that for the initial data u_0 constructed in §3 no additional zeros are created by the flow, we as yet do not have a proof of this. In any event, the existence of such an example is slightly surprising, since, intuitively, one might guess that the creation of additional zeroes would violate the dissipation of energy (2.15).

Since the effect does not require $\varepsilon << 1$, for simplicity we will let u denote the solution to (1.1)–(1.3) with $\varepsilon = 1$. We take $\Omega \subset \mathbf{R}^2$ to be any bounded domain containing the origin and g to be arbitrary smooth boundary data. Then we require the initial data $u_0 = (u_0^{(1)}, u_0^{(2)})$ to be any smooth function compatible with g on $\partial\Omega$, possessing k zeros q_i, $i = 1, 2, \ldots k$, $q_i \neq 0$ with

$$\sum_{i=1}^k \deg(u_0, \partial B_r(q_i)) = \deg(g, \partial\Omega) \text{ for all small } r,$$

and such that u_0 satisfies

$$
\begin{aligned}
u_0^{(1)}(x_1, x_2) &= x_2 - \delta - \frac{1}{2}x_1^2 + (x_2 - \delta)^2, \\
u_0^{(2)}(x_1, x_2) &= x_2 + \delta + x_1^2 - \frac{1}{2}(x_2 + \delta)^2
\end{aligned} \tag{4.1}
$$

in some ball $B_{r_0}(0)$, where $\delta > 0$. The relevant features of this choice of data are that

$$u_0^{(1)}(0, \delta) = 0 = u_0^{(2)}(0, -\delta), \tag{4.2}$$

while for all sufficiently small r_0 we have

$$\{u_0^{(1)} = 0\} \cap \{u_0^{(2)} = 0\} \cap B_{r_0}(0) = \emptyset \tag{4.3}$$

(see Fig. 1).

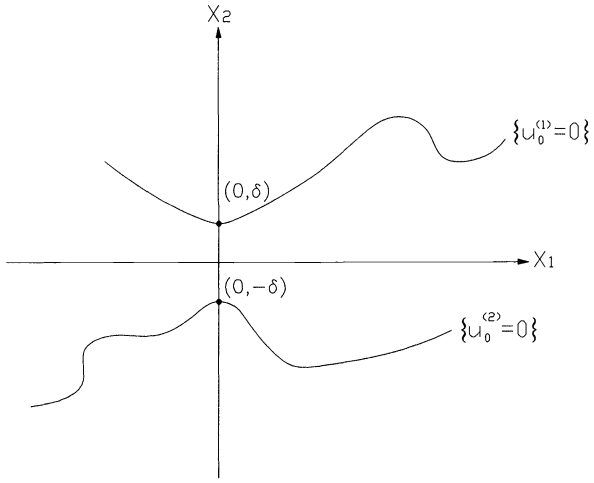


FIG. 1. *The zero sets of the two components of the initial data in a neighborhood of the origin.*

Henceforth we fix r_0 such that (4.1) and (4.3) hold. Note that such a choice can be made so that (4.3) is valid for all small δ. We will then show the following proposition.

PROPOSITION 4.1. *Let u solve* (1.1)–(1.3) *with $\varepsilon = 1$, where Ω, g, and u_0 are as described above. In particular, we assume* (4.1) *and* (4.3) *hold for some $r_0 > 0$. Then there exists $\delta_0 > 0$ such that for all positive $\delta \leq \delta_0$, the corresponding solution u develops two zeroes $q_1(t)$ and $q_2(t)$ in $B_{r_0}(0)$ after some finite time. Immediately after their emergence, these zeroes will satisfy*

$$\deg(u, \partial B_r(q_i(t))) = \pm 1, \quad i = 1, 2$$

for all r sufficiently small.

Remark. We do not claim that these zeroes will necessarily persist. Indeed, one would typically expect that they will annihilate each other shortly after the time of their creation.

Proof. The argument consists of showing that the vertical component of the velocity of the level set $\{u^{(1)} = 0\}$ (resp., $\{u^{(2)} = 0\}$) is negative (resp., positive) and

bounded away from zero uniformly in δ for all small times in a neighborhood of the origin. In light of (4.2), this implies that the two zero sets must collide, resulting in the creation of zeroes of opposite degree.

To this end let $z_1 = z_1(x_1, t)$ denote the local graph of the zero set of $u^{(1)}$ which includes the point $(0, \delta)$ at time $t = 0$, i.e., $z_1(0,0) = \delta$. Similarly, we let $z_2 = z_2(x_1, t)$ denote the local graph of the zero set of $u^{(2)}$ such that $z_2(0,0) = -\delta$. Since u_0 is taken to be smooth, one readily obtains local C^2 bounds on the solution u in the cylinder $B_{r_0}(0) \times [0, T]$ for any $T > 0$. Such bounds and the condition

$$(u_0^{(1)})_{x_2}(0, \delta) = 1 = (u_0^{(2)})_{x_2}(0, -\delta) \tag{4.4}$$

imply that z_1 and z_2 are well defined and smooth in the cylinder $B_{r_0} \times [0, T^*]$ for some T^* small (but not dependent on δ since $\|u\|_{C^2}$ is independent of δ).

To obtain the initial velocity of the graphs z_1 and z_2, we differentiate the equations

$$u^{(i)}(x_1, z_i(x_1, t), t) = 0, \quad i = 1, 2$$

with respect to t and use (1.1) (with $\varepsilon = 1$) to find

$$\begin{aligned}(z_i)_t u_{x_2}^{(i)} = -(u^{(i)})_t &= -\Delta u^{(i)} - u^{(i)}(1 - |u|^2) \\ &= -\Delta u^{(i)}, \quad i = 1, 2,\end{aligned} \tag{4.5}$$

where the nonlinear term vanishes, since we are evaluating $u^{(i)}$ along its zero set. Using the fact that u is C^2 in $B_{r_0}(0) \times [0, T^*]$, we may take the limit in (4.5) as $t \to 0$ and use (4.1) to find

$$(z_1)_t(0,0) = -1 \quad \text{and} \quad (z_2)_t(0,0) = 1.$$

Hence there exists a positive time $\tilde{T} \le T^*$ and a positive number A, both depending on $\|u\|_{C^2}$ but independent of δ, such that

$$\begin{aligned}(z_1)_t(x_1, t) &\le -\frac{1}{2} \text{ for } |x_1| \le A,\ 0 \le t \le \tilde{T}, \\ (z_2)_t(x_1, t) &\ge \frac{1}{2} \text{ for } |x_1| \le A,\ 0 \le t \le \tilde{T}.\end{aligned}$$

Since $z_1(0,0) = \delta$ and $z_2(0,0) = -\delta$, we conclude that there exists a $\delta_0 > 0$ such that for all positive $\delta \le \delta_0$, the two graphs cross and the corresponding solution to (1.1)–(1.3) develops two zeroes (of degree ± 1) after a short time in a neighborhood of the origin. □

REFERENCES

[BBH1] F. Bethuel, H. Brezis, and F. Helein, *Limite singuliere pour la minimisation de fonctionnelles du type Ginzburg-Landau*, C.R. Acad. Sci. Paris Ser. I Math., 314 (1992), pp. 891–895.

[BBH2] F. Bethuel, H. Brezis, and F. Helein, *Tourbillons de Ginzburg-Landau et energie renormalisee*, C.R. Acad. Sci. Paris, 317 (1993), pp. 165–171.

[BCP] P. Bauman, N. Carlson, and D. Phillips, *On the zeroes of solutions to Ginzburg–Landau type systems*, SIAM J. Math. Anal., 24 (1993), pp. 1283–1293.

[BCPS] P. Bauman, C.-N. Chen, D. Phillips, and P. Sternberg, *Annihilation of vortices for the Ginzburg–Landau heat flow*, European J. Appl. Math., to appear.

[BKP] E. BODENSCHATZ, W. PESCH, AND L. KRAMER, *Structure and dynamics of dislocations in anisotropic pattern-forming systems*, Physica D, 32 (1988), pp. 135–145.

[FP] P. FIFE AND L. A. PELETIER, *On the location of defects in stationary solutions of the Ginzburg–Landau equation in* $\mathbf{R}^2$, preprint.

[GK] Y. GIGA AND R. V. KOHN, *Characterizing blowup using similarity variables*, Indiana Univ. Math. J., 36 (1987), pp. 1–40.

[GT] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[K] Y. KURAMOTO, *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, New York, Berlin, Heidelberg, 1984.

[LSU] O. LADYŽENSKAJA, V. SOLONNIKOV, AND N. URALĆEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[N1] J. NEU, *Evolution and creation of vortices and domain walls*, Department of Mathematics, University of California, Berkeley, lectures notes (unpublished).

[N2] ———, *Vortices in complex scalar fields*, Physica D, 43 (1990), pp. 385–406.

[PRo] L. PISMEN AND J. D. RODRIGUEZ, *Mobility of singularities in the dissipative Ginzburg–Landau equation*, Phys. Rev. A, 42 (1990), pp. 2471–2474.

[PR1] L. PISMEN AND J. RUBINSTEIN, *Dynamics of defects*, in Nematics: Mathematical and Physical Aspects, J.M. Coron et al., eds., Kluwer Academic Publishers, Norwell, MA, 1991.

[PR2] ———, *Dynamics of disclinations in liquid crystals*, Quart. Appl. Math., 50 (1992), pp. 535–545.

[PZM] Y. POMEAU, S. ZALESKI, AND P. MANNEVILLE, *Disclination motion in cellular structures*, Phys. Rev. A, 27 (1983), pp. 2710–2726.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1467-1472, November 1995

005

A UNIQUENESS RESULT FOR A GENERALIZED RADON TRANSFORM*

B. L. FRIDMAN†

Abstract. There exist five families of Lipschitz curves on the unit square such that any continuous function is uniquely defined by the values of its integral (properly defined) along these curves. We present this uniqueness result as a consequence of the Kolmogorov superposition theorem.

Key words. radon transform, spreads, uniqueness, superposition

AMS subject classifications. 65R10, 44A12, 92C55, 26B40

1. Introduction. The main intention of this article is to demonstrate a connection of results in linear superpositions of functions and the uniqueness problem for a generalized Radon transform. We will actually present one such relation, namely, a consequence of the Kolmogorov superposition theorem. We will start with a description of a general problem.

Suppose that for each $\alpha \in J$ (J is the set of indices) there is a family (also called a *spread*) Ω_α of nonintersecting submanifolds $\Gamma_t^\alpha, t \in T = T_\alpha$, of some manifold $M = \bigcup_{t\in T} \Gamma_t^\alpha$, and for each $\Gamma = \Gamma_t^\alpha$ there is a measure $d\mu_\Gamma$ such that one can introduce a generalized Radon transform by

$$Rf(\Gamma) = \int_\Gamma f d\mu_\Gamma \tag{1}$$

for an integrable function f. Typically, Γ_t^α are smooth hypersurfaces, the parameter set T is a one-dimensional interval, $M \subseteq \mathbf{R^n}$, and f is a continuous function with compact support. The general uniqueness (or invertibility of R) problem we consider here is as follows: If $Rf(\Gamma_t^\alpha) = Rf(\alpha, t) = 0$ for all α, t then $f = 0$. Similar problems have been addressed in a number of papers (see [1]–[6], [9]–[10], [12]–[16]).

The question we discuss here is the following one: How many families (the cardinality of J) does it take to assure the uniqueness (invertibility) of R? Intuitively, it seems very likely that if the number of different families is infinite, the uniqueness takes place, and if the number of families is finite, the uniqueness does not hold. In many cases that have been considered, this assertion is supported; however, one should recall the result of Boman [1] showing that the uniqueness is not necessarily assured in case Γ_t^α are straight lines in $\mathbf{R^2}$, $\alpha \in J = [0, 2\pi]$ is infinite, and the measure is C^∞ and positive.

Here we consider the case of a *finite* number of families. Is it possible to find a finite system of families such that uniqueness holds for continuous functions? Unexpectedly, the answer to this question in a typical case is positive. The main result of this paper is the following one: There exist five families of Lipschitz curves on the unit square in $\mathbf{R^2}$ and such a measure on each of these curves that the uniqueness property holds for all continuous functions.

For the special case we are describing we have $\Gamma = \Gamma_t^\alpha$ as curves on the closure of the unit square $\Delta \subset \mathbf{R^2}$; we also use the notation $I_\Gamma(f)$ instead of $Rf(\Gamma)$ for

* Received by the editors June 30, 1993; accepted for publication (in revised form) March 3, 1994.
† Mathematics Department, Wichita State University, Wichita, Kansas 67260-0033 (fridman@twsuvm.uc.twsu.edu).

the corresponding integral (1). We also make one more remark. In many cases it is desirable to consider *natural* measures $d\mu_\Gamma$ such that the Fubini theorem holds: the "double" integral of f over M could be presented as a repeated integral, that is, an integral of $I_{\Gamma_t^\alpha}(f)$ over T^α with some measure $d\mu_\alpha(t)$. The measure we introduce below is such a natural measure (see (3)).

As stated above this result is going to be a consequence of the well-known theorem by Kolmogorov, that is, the theorem that presented a solution of the 13th problem of Hilbert. We use the notation Δ for the closure of the unit square: $\Delta = \{(x,y)|0 \leq x,y \leq 1\}$.

THEOREM A (superposition theorem of Kolmogorov). *There exist five functions* $\Phi_i(x,y) = \varphi_i(x) + \psi_i(y)$, $i = 1,\dots,5$, *such that for any* $f \in C(\Delta)$ *there exist continuous functions* $\chi_i(t)$, *so that*

$$f(x,y) = \sum_{i=1}^{5} \chi_i(\varphi_i(x) + \psi_i(y)).$$

The functions $\varphi_i(x), \psi_i(y)$ can be chosen strictly increasing and Lip1 functions. For proof see [11]; the choice of Lipschitz functions was proved in [7].

2. Definitions and basic results. A reasonable way to introduce a family of curves is to consider them as level sets of a function. Following a suggestion of Ehrenpreis, such a family will be called a *spread* and the corresponding function will be a *spread function.* We present the definition in two steps.

1) Let D be a domain in $\mathbf{R^2}$ and $\Phi \in C(\overline{D})$. Consider $\Gamma_t = \{(x,y)|\Phi(x,y) = t\} \cap \overline{D}$, the level curve of Φ, and $[a,b] = \Phi(\overline{D})$, the range of Φ. Clearly, $\Gamma_t \cap \Gamma_\tau = \emptyset$ if $t \neq \tau$ and $\overline{D} = \bigcup_{t\in[a,b]} \Gamma_t$. We call the set $\Omega = \{\Gamma_t \mid t \in [a,b]\}$ a *spread* on $\overline{D}$ and we call Φ the *spread function.*

2) A spread Ω on $\overline{D}$ generated by a spread function Φ is called a *proper spread* if the following holds. There exists a homeomorphism $\phi : \overline{D} \to \overline{U} = \phi(\overline{D})$ such that the set $\{\phi(\Gamma_c)|\Gamma_c \in \Omega\}$ is a set of straight parallel lines on $\overline{U}$: $\phi(\Gamma_t) \| \phi(\Gamma_\tau)$ for any $t,\tau \in [a,b]$.

All the results that follow hold for many domains in R^2. Without any significant loss of generality and for simplicity of exposition we consider our domain to be the unit square Δ.

Suppose we have a proper spread Ω of curves on Δ generated by a continuous function Φ and $\Gamma = \Gamma_t \in \Omega$. Let $f \in C(\Gamma)$. Then f can be extended to a continuous function on Δ, which we will still denote by f. We now introduce $I_\Gamma(f)$, the integral of f over Γ. Let $\Gamma(\epsilon) = \{(x,y) \in \Delta|\Phi(x,y) \in (t-\epsilon, t+\epsilon)\}$. Then we define

$$I_\Gamma(f) = \lim_{\epsilon\to 0} \frac{\int_{\Gamma(\epsilon)} f dA}{\mathrm{mes}_2(\Gamma(\epsilon))}, \tag{2}$$

where dA is the area element and mes_2 is the Lebesgue area.

Example. One can check the following. If $\Phi \in C^1(\Delta)$ and $\mathrm{grad}\Phi \neq 0$ on $\Gamma = \Gamma_t$ with positive length, then $I_\Gamma(f) = \int_\Gamma f / \| \mathrm{grad}\Phi \| ds / l_g(\Gamma)$, where ds is the element of length and $l_g(\Gamma)$ is the weighted length of Γ; $l_g(\Gamma) = \int_\Gamma 1/\| \mathrm{grad}\Phi \| ds$.

LEMMA 1. 1. *Let* Γ *from* (2) *be fixed. If* $I_\Gamma(f)$ *exists for some continuous extension of* f *from* Γ *to* Δ, *then it exists and has the same value for any such continuation of* f *to* Δ. *In this case* $| I_\Gamma(f) | \leq \max_{x\in\Gamma} |f(x)|$.

2. *The set* $E(\Gamma)$ *of continuous functions on* Γ *for which* $I_\Gamma(f)$ *exists forms a closed linear subspace in* $C(\Gamma)$. $I_\Gamma(f)$ *is a linear bounded functional on* $E(\Gamma)$.

3. *If* $f = c = \text{const.}$ *on* Γ *then* $I_\Gamma(f) = c$.

4. *If* $\chi(t) \in C[a,b]$ *then* $I_{\Gamma_t}(\chi(\Phi(x,y)) = \chi(t)$.

Proof. First one can check 3. It follows from the fact that the extension is continuous and also from the mean value theorem for the Lebesgue integral on bounded functions.

One can now check 1. The first assertion follows from the following observation. The difference between two extensions of f is zero on Γ and, therefore, the integral of the difference will always exist and be equal to zero. One can prove the inequality by using the standard inequalities for the integral of a bounded function.

Statement 2 follows from the linear property of the Lebesgue integral and the already proved inequality in 1.

Statement 4 follows from 3. □

Denote by Φ_* the push-forward induced by the function $\Phi : \Delta \to \Phi(\Delta) = [a,b] = J$. (The push-forward takes measures on Δ to measures on J.) We now introduce a countably additive measure μ on J by the formula $\mu(dt) = \Phi_*(dA)$ (that is, $\mu(T) = \text{mes}_2(\Phi^{-1}(T))$ for $T \subseteq J$). One can see that μ is a countably additive measure on J and all Borel sets on J are μ-measurable. Let $f \in C(\Delta)$. We may now consider a measure ν_f given by $\nu_f(dt) = \Phi_*(f dA)$. Clearly, ν_f must be absolutely continuous with respect to μ. Therefore, by the Radon–Nikodim theorem there exists $g \in L_1(\mu)$ such that $\nu_f = g\mu$, where

$$g(t) = \lim_{\epsilon \to 0} \frac{\nu_f(t-\epsilon, t+\epsilon))}{\mu(t-\epsilon, t+\epsilon)} = \lim_{\epsilon \to 0} \frac{\int_{\Gamma_t(\epsilon)} f dA}{\text{mes}_2(\Gamma_t(\epsilon))}$$

and $g(t)$ exists almost everywhere with respect to the μ -measure (almost everywhere (a.e.) μ) on J. If one compares this formula with the definition (2) of $I_\Gamma(f)$, one has the proof of the first part of the following lemma.

LEMMA 2. 1. $g(t) = I_{\Gamma_t}(f)$ *a.e.* μ *and, therefore, the last integral is uniquely defined for a fixed* f *a.e.* μ *on* J. *In other words, the union of all level curves* Γ *for which the limit in* (2) *does not exist, forms a set of Lebesgue area zero.*

2. *The following formula holds:*

$$\int_\Delta f dA = \int_a^b I_{\Gamma t}(f) \mu(dt). \tag{3}$$

Proof. The proof of the last formula follows from the following observation. Considering measures as linear functionals on the space of continuous functions, we know that the push-forward Φ_* is the adjoint of the pull-back $\Phi^* : u \longmapsto u \circ \Phi$ for $u \in C(J)$, that is,

$$\langle v, u \circ \Phi \rangle = \langle v, \Phi^*(u) \rangle = \langle \Phi_*(v), u \rangle.$$

Taking $u = 1$ and $v = f dA$, we obtain (3). □

We will say that a certain property holds for *almost every* Γ if the union S of all Γ for which this property fails has Lebesgue area zero: $mes_2(S) = 0$. Obviously, this definition does not depend on the spread function generating the set of curves; it depends on the set of curves (the spread) only.

The space $E(\Gamma)$ (see Lemma 1) is nonempty and $I_\Gamma(f)$ is a bounded linear functional on this space. Obviously, $E(\Gamma) \subseteq C(\Gamma)$.

LEMMA 3. *$E(\Gamma) = C(\Gamma)$ for almost every Γ (a.e. Γ in the above-defined meaning).*

Proof. Let set S be the union of all such Γ for which $E(\Gamma) \neq C(\Gamma)$. Consider all polynomials of two variables with rational coefficients. They form a countable set $\{P_1, P_2, \ldots\}$. Let S_n be the union of all such Γ for which $I_\Gamma(P_n)$ does not exist. By Lemma 2, $\text{mes}_2(S_n) = 0$. Since every continuous function can be approximated uniformly on Δ by this set of polynomials, $S \subseteq \bigcup_{n=1}^{\infty} S_n$ and, therefore, $\text{mes}_2(S) = 0$. □

LEMMA 4. *If $f \in C(\Delta)$ and $\chi(t) \in C[a,b]$ then for almost all (relative to the μ measure) $t \in [a,b]$,*

$$I_{\Gamma_t}\{f(x,y)\chi(\Phi(x,y)\} = I_{\Gamma_t}(f)\chi(t). \tag{4}$$

Proof. The proof follows from the linear property of the introduced integral and the uniqueness of the introduced integral from Lemma 2. □

3. The main statement. Now we are ready to prove the following theorem.

THEOREM B. *There exist five proper spreads Ω_i, $i = 1, \ldots, 5$ of curves on Δ such that if $f \in C(\Delta)$ and $I_\Gamma(f) = 0$ for almost every curve Γ in any Ω_i, then $f \equiv 0$.*

Remark. Almost every Γ has the same meaning as before: all Γ with the exception of a set of curves whose union has zero area in Δ.

Proof. Consider $\Phi_i(x,y) = \phi_i(x) + \psi_i(y)$ from Theorem A. As noted above all these functions can be chosen to be strictly increasing and satisfy the Lipschitz condition. We can also assume that $\phi_i(\Delta) = [0, q_i], \psi_i(\Delta) = [0, p_i]$, so $\Phi(\Delta) = [0, q_i + p_i]$. Now consider $\Gamma_t^i = \{(x,y) \in \Delta | \Phi_i(x,y) = t\}$.

1. First we prove that $\Omega_i = \{\Gamma_t^i\}$ is a proper spread. Consider the system

$$t = \phi_i(x) + \psi_i(y)$$
$$u = \phi_i(x) - \psi_i(y)$$

One can show, using the monotonicity of the functions involved, that the above system provides a homeomorphism of Δ onto the rectangle $0 \leq t+u \leq 2q_i, 0 \leq t-u \leq 2p_i$ and, therefore, Ω_i is in one-to-one correspondence with parallel lines $t = \text{const}$.

2. Let $f \in C(\Delta)$ be a function for which $I_{\Gamma_t^i}(f) = 0$ for almost all Γ_t^i. According to Theorem A, $f(x,y) = \sum_{i=1}^{5} \chi_i(\Phi_i(x,y))$. Now consider (we use (3) and (4))

$$\begin{aligned}
\iint_\Delta f^2(x,y)dA &= \iint_\Delta f(x,y) \sum_{i=1}^{5} \chi_i(\Phi_i(x,y))dA \\
&= \sum_{i=1}^{5} \int_0^{q_i+p_i} I_{\Gamma_t^i}\{f(x,y)\chi_i(\Phi_i(x,y))\}\mu_i(dt) \\
&= \sum_{i=1}^{5} \int_0^{q_i+p_i} I_{\Gamma_t^i}(f)\chi_i(t)\mu_i(dt) = 0.
\end{aligned}$$

Therefore $f \equiv 0$. □

Remark 1. The constructed spreads Ω_i consist of Lipschitz rectifiable curves. The $I_\Gamma(f)$ is not necessarily absolutely continuous for every Γ.

Problem. Will the statement of Theorem B hold if $I_\Gamma(f) = \int_\Gamma f ds$, where ds is the element of the arc length?

Remark 2. Theorem A (see [11]) was proved for a general case in R^n for all n. Similarly, one can prove a statement analogous to Theorem B in case $n > 2$, replacing curves by surfaces.

Remark 3. One can check that the construction in (2) and most constructions that followed can be done for $f \in L_1(\Delta)$ instead of a continuous function. This would lead to the following theorem.

THEOREM B′. *There exist five proper spreads* Ω_i , $i = 1, \ldots, 5$ *of curves on* Δ *such that if* $f \in L_1(\Delta)$ *and* $I_\Gamma(f) = 0$ *for almost every curve* Γ *in any* $\Omega_i, i = 1, \ldots, 5$, *then* $f = 0$.

To prove this, one can repeat the proof presented above and consider $\iint_\Delta fgdA$, $g \in C(\Delta)$. Then one can proceed as above to prove that this integral is zero and thus f is orthogonal to any continuous function. Therefore f is equal to zero in L_1.

The following statement shows the critical difference in the outcome if we replace the Lipschitz condition for functions Φ_i generating our families of curves by the requirement that these functions be continuously differentiable.

THEOREM C. *Let spreads* Ω_i , $i = 1, \ldots, s$ *of curves on* Δ *be generated by spread functions* $\Phi_i \in C^1(\Delta)$. *Then there exists a function* $f \in L_2(\Delta), f \neq 0$, *such that* $I_\Gamma(f) = 0$ *for almost every curve* Γ *in any* $\Omega_i, i = 1, \ldots, s$.

Proof. The proof follows from the result in [8, Thm. 4] that proves that the set $G = \{g \mid g(x,y) = \sum_{i=1}^{s} \chi_i(\Phi_i(x,y)), \chi_i \in C\}$ is nowhere dense in $L_2(\Delta)$. Now taking $f \in L_2(\Delta)$, $f \perp G$, one can check, using the same idea as above, that $I_\Gamma(f) = 0$ for almost every Γ in any Ω_i. □

Acknowledgment. I am very grateful to Professor Peter Kuchment for valuable discussions and encouragement. I am also most thankful to the referee for useful suggestions that improved the exposition of the paper.

REFERENCES

[1] J. BOMAN, *An example of non-uniqueness for a generalized Radon transform*, J. Analyse Math., 61 (1993), pp. 395–399.

[2] ———, *Uniqueness theorems for generalized Radon transforms*, in Constructive Theory of Functions '84, Sofia, 1984, pp. 173–176.

[3] ———, *Helgason's support theorem for Radon transforms—a new proof and a generalization*, in Mathematical Methods in Tomography, Proceedings of a conference held at Oberwolfach 1990, Lecture Notes in Mathematics 1497, Springer-Verlag, Berlin, New York, 1991.

[4] J. BOMAN AND E. T. QUINTO, *Support theorems for real-analytic Radon transforms*, Duke Math. J., 55 (1987), pp. 943–948.

[5] ———, *Support theorems for Radon transform on real-analytic line complexes in three-space*, Trans. Amer. Math. Soc., 335 (1993), pp. 877–890.

[6] D. V. FINCH, *Uniqueness of the attenuated X-ray transform in the physical range*, Inverse Problems, 2 (1986), pp. 197–203.

[7] B. L. FRIDMAN, *Improvement in the smoothness of functions in the Kolmogorov superposition theorem*, Dokl. Akad. Nauk SSSR, 177 (1967), pp. 1019–1022; Soviet Math. Dokl., 8 (1967), pp. 1550–1553.

[8] ———, *An estimate of the dimension of the null spaces of linear superpositions*, Mat. Sb., 82 (1970), pp. 111–125; Math. USSR-Sb., 11 (1970), pp. 101–114.

[9] A. GREENLEAF AND G. UHLMANN, *Non-local inversion formulas for the X-ray transform*, Duke Math. J., 58 (1989), pp. 205–240.

[10] S. HELGASON, *Groups and Geometric Analysis*, Academic Press, New York, 1984.

[11] A. N. KOLMOGOROV, *On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition*, Dokl. Akad. Nauk SSSR, 114 (1957), pp. 953–956; Amer. Math. Soc. Transl., 28 (1963), pp. 55–59.

[12] A. MARKOE AND E. T. QUINTO, *An elementary proof of local invertibility for generalized and attenuated Radon transforms*, SIAM J. Math. Anal., 16 (1985), pp. 1114–1119.

[13] R. G. MUKHOMETOV, *The problem of recovery of a two-dimensional Riemannian metric and integral geometry*, Dokl. Akad. Nauk SSSR, 232 (1977), pp. 32–35; Soviet Math. Dokl., 18 (1977), pp. 27–31.

[14] ———, *A problem of reconstructing a Riemannian metric*, Sibirsk. Mat. Zh., 22 (1981), pp. 119–135; Siberian Math. J., 22 (1981), pp. 420–433.

[15] E. T. QUINTO, *The invertibility of rotation invariant Radon transforms*, J. Math. Anal. Appl., 91 (1983), pp. 510–522.

[16] L. ZALCMAN, *Uniqueness and nonuniqueness for the Radon transform*, Bull. London Math. Soc., 14 (1982), pp. 241–245.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1473–1487, November 1995

006

MONOTONICITY AND INVERTIBILITY OF COEFFICIENT-TO-DATA MAPPINGS FOR PARABOLIC INVERSE PROBLEMS*

PAUL DUCHATEAU†

Abstract. This paper considers the coefficient-to-data mappings associated with unknown coefficient inverse problems for nonlinear parabolic partial differential equations. Integral identities are derived that show the coefficient-to-data mapping is monotone and invertible in the case of a single unknown coefficient and the mapping is invertible in the case of simultaneous determination of two unknown coefficients.

Key words. inverse problems, parabolic equations

AMS subject classifications. 35R30, 35Q99, 35K15

Introduction. The determination of unknown coefficients in parabolic partial differential equations from overspecified data measured on the boundary is a problem of some importance in applied mathematics. Such so-called inverse problems arise naturally, for example, in modeling nonlinear diffusion and flow in porous media. Direct measurement of the quantities represented by the unknown coefficients often requires very difficult physical experiments. The point of the inverse problems is to replace a difficult physical experiment by a mathematical problem for which the input is easy to measure. The ease of measurement requirement suggests that the data be measured on the boundary.

Heretofore the method of "output least squares" has been a popular approach to solving unknown coefficient inverse problems [1], [2]. Here the inverse problem is replaced by an optimization problem designed to select coefficients that produce a solution that best matches some measured output. Examples show that the choice of output must be made with care [3]. The flaw in output least squares seems to lie in the fact that there is no way of proving that the solution to the optimization problem is a solution to the original inverse problem.

Previous analyses have succeeded in showing that the solutions of certain inverse problems are unique if they exist [4]–[6], but until now it has been unclear whether the inverse problems are in fact solvable. Results obtained here suggest the inverse problems are solvable and the coefficients can be computed from rather simple algorithms based on very explicit representations for the associated coefficient-to-data mappings.

This paper is organized as follows. In §1 the so-called direct problem is formulated and the properties of the measured output are deduced from the properties of the admissible inputs. Integral identities that relate changes in equation coefficients to changes in measured output are derived in §2. These identities are, in effect, explicit representations of the coefficient-to-data mappings, and the final two sections are devoted to showing how they may be used to constructively solve parabolic inverse problems.

1. The direct problem.

Suppose that

(i) $C \in \mathbb{C}(0,\infty)$ with $0 < C_0 \le C(u) \le C_1$ for $0 < u$,

* Received by the editors November 30, 1993; accepted for publication (in revised form) March 24, 1994.

† Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523.

(ii) $K \in \mathbb{C}^1(0,\infty)$ with $0 < K_0 \le K(u) \le K_1$ for $0 < u$.

Let

$$a(u) = \int_{u_0}^{u} C(s)\,ds \quad \text{and} \quad b(u) = \int_{u_0}^{u} K(s)\,ds$$

so that $a'(u) = C(u)$, $b'(u) = K(u)$, and $a(u_0) = b(u_0) = 0$. Then C and K will be said to be *admissible coefficients* for the initial boundary value problem for the unknown function $u(x,t)$:

$$\begin{aligned} \partial_t a(u) &= \partial_{xx} b(u) \quad \text{for } 0 < x < 1,\ 0 < t < T, \\ u(x,0) &= u_0 \quad \text{for } 0 < x < 1, \\ \partial_x u(0,t) = 0 \text{ and } u(1,t) &= f(t) \quad \text{for } 0 < t < T. \end{aligned} \tag{1.1}$$

For each pair of admissable coefficients and data $u_0 \ge 0$, $f(t) \in \mathbb{C}[0,T]$, there exists a unique smooth solution $u(x,t)$ for (1.1). One refers to (1.1) as the direct problem as opposed to the inverse problem in which one seeks to determine the coefficients $a(u)$ and $b(u)$ from overspecified data measured on the boundary. For the purpose of finding unknown coefficients, two overspecifications are considered:

$$h(t) = u(0,t) \quad \text{and} \quad g(t) = \partial_x b(u(1,t)) = K(f(t))\partial_x u(1,t). \tag{1.2}$$

The main purpose of this paper is to show that under appropriate assumptions on the input data, $f(t)$, the coefficient-to-data mapping for each of these overspecifications is an invertible mapping. The input data, $f(t)$, will be termed admissible if $f(0) = u_0 \ge 0$ and $f'(t) > 0$ for $t > 0$. Assuming that the coefficients and input are admissible, consequent properties can be deduced for the solution $u(x,t)$ of the direct problem and for the overspecified functions g and h.

LEMMA 1.1. *Let $u(x,t)$ denote the solution of the direct problem corresponding to admissible coefficients C and K and admissible data $f(t)$. Then for each τ, $0 < \tau \le T$, $u_0 < u(x,t) < f(\tau)$ for $0 < x < 1$, and $0 < t \le \tau$. In addition, $g(t) = \partial_x b(u)(1,t) = K(f(t))\partial_x u(1,t) > 0$ for $t > 0$.*

Proof. For any τ, $0 < \tau \le T$, one has, for admissible $f(t)$, $u_0 \le f(t) < f(\tau)$ for $0 \le t < \tau$. Let M_τ and m_τ, respectively, denote the maximum and minimum values of u on the parabolic boundary of $Q_\tau = (0,1) \times (0,\tau)$. Then the strong maximum principle for parabolic equations implies $m_\tau < u(x,t) < M_\tau$ on the open domain Q_τ. Neither m_τ nor M_τ occurs on $x = 0$ because $\partial_x u(0,t) = 0$. Then m_τ occurs at $t = 0$ and M_τ occurs at $x = 1$; i.e., $u(x,0) = u_0 = m_\tau$ and $u(1,\tau) = f(\tau) = M_\tau$. In particular, $u(1-h,\tau) < u(1,\tau) = M_\tau$ for all $h > 0$. This implies $\partial_x u(1,t)$ is positive since

$$\partial_x u(1,\tau) = \lim_{h \to 0+} \frac{u(1,\tau) - u(1-h,\tau)}{h} > 0 \quad \text{for all } \tau > 0.$$

LEMMA 1.2. *Let $u(x,t)$ solve the direct problem for admissible coefficients and input. Then $\partial_x u(x,t) \ge 0$ on the open set $Q_T = \{0 < x < 1, 0 < t < T\}$ and there is no positive measure subset of Q_T where $\partial_x u(x,t) = 0$.*

Proof. For an arbitrary smooth function $\varphi(x,t)$

$$\int\int_{Q_T} [\partial_t a(u) - \partial_{xx} b(u)]\partial_x \varphi \, dx\, dt = 0$$

and integration by parts leads to the result

$$\int\int_{Q_T} \partial_x u[C(u)\partial_t\varphi + K(u)\partial_{xx}\varphi]\,dx\,dt = \int_0^T \partial_x b(u)\partial_x\varphi\Big|_{x=0}^{x=1} dt + \int_0^T a(u)\partial_t\varphi\Big|_{x=0}^{x=1} dt - \int_0^1 a(u)\partial_x\varphi\Big|_{t=0}^{t=T} dx. \tag{1.3}$$

If $\varphi(x,t)$ is chosen to be the solution of *adjoint problem* 1:

$$\begin{aligned} C(u)\partial_t\varphi + K(u)\partial_{xx}\varphi &= F(x,t) \quad \text{in } U_T, \\ \varphi(x,T) &= 0, \quad 0 < x < 1, \\ \varphi(0,t) = 0 \text{ and } \varphi(1,t) &= 0, \quad 0 < t < T, \end{aligned} \tag{1.4}$$

then $\partial_x\varphi(x,T) = 0$, $\partial_t\varphi(0,t) = \partial_t\varphi(1,t) = 0$, and (1.3) reduces to

$$\int\int_{Q_T} \partial_x u(x,t)F(x,t)\,dx\,dt = \int_0^T K(u)\partial_x u(1,t)\partial_x\varphi(1,t)\,dt. \tag{1.5}$$

The maximum principle applied to adjoint problem 1 shows that if $F(x,t) \geq 0$ on Q_T, then $\varphi < 0$ on Q_T and, since $\varphi(1,t) = 0$, it follows that $\partial_x\varphi(1,t) > 0$. Lemma 1 and the hypotheses on $f(t)$ imply $g(t) = K(f(t))\partial_x u(1,t) > 0$ for $t > 0$. Then the right side of (1.5) is strictly positive. $F(x,t)$ has been assumed to be nonnegative but is otherwise arbitrary, hence it follows from (1.5) that $\partial_x u(x,t)$ is nonnegative in Q_T. Similarly, the existence of a positive area subset in Q_T where $\partial_x u(x,t)$ vanishes leads to a contradiction with (1.5).

LEMMA 1.3. *Let $u(x,t)$ solve the direct problem for admissible coefficients and input. Then $\partial_t u(x,t) \geq 0$ on Q_T and there can be no set of positive measure in Q_T where $\partial_t u$ vanishes.*

Proof. For an arbitrary smooth function $\varphi(x,t)$

$$\int\int_{Q_T} [\partial_t a(u) - \partial_{xx} b(u)]\partial_t\varphi\,dx\,dt = 0.$$

Therefore, integration by parts shows that

$$\int\int_{Q_T} \partial_t u[C(u)\partial_t\varphi + K(u)\partial_{xx}\varphi]\,dx\,dt = \int_0^T K(u)[\partial_t u\partial_x\varphi + \partial_x u\partial_t\varphi]_{x=0}^{x=1}\,dt - \int_0^1 K(u)\partial_x u\partial_x\varphi\Big|_{t=0}^{t=T} dx. \tag{1.6}$$

If $\varphi(x,t)$ is chosen to be the solution of *adjoint problem* 2:

$$\begin{aligned} C(u)\partial_t\varphi + K(u)\partial_{xx}\varphi &= F(x,t) \quad \text{in } Q_T, \\ \varphi(x,T) &= 0, \quad 0 < x < 1, \\ \partial_x\varphi(0,t) = 0 \text{ and } \varphi(1,t) &= 0, \quad 0 < t < T, \end{aligned} \tag{1.7}$$

then $\partial_t\varphi(1,t) = 0$, $\partial_t u(1,t) = f'(t)$, and since $\partial_x\varphi(0,t) = 0$ and $\partial_x u(0,t) = 0$, the integral identity (1.6) reduces to

$$\int\int_{Q_T} \partial_t u(x,t)F(x,t)\,dx\,dt = \int_0^T K(u)f'(t)\partial_x\varphi(1,t)\,dt. \tag{1.8}$$

The maximum principle applied to adjoint problem 2 shows that $F \geq 0$ in Q_T implies $\varphi < 0$ in Q_T (i.e., replace (1.7) with an extended problem on $(-1,1) \times (0,T)$ with $\varphi(-1,t) = 0$. The maximum principle applies to the extended problem exactly as it did for problem (1.4).). Then $\varphi < 0$ in Q_T implies that $\partial_x \varphi(1,t) > 0$ and this together with the hypotheses on $f(t)$ ensure that the right side of (1.8) is strictly positive. $F(x,t)$ has been assumed to be nonnegative but is otherwise arbitrary, thus it follows from (1.8) that $\partial_t u \geq 0$ in Q_T. Moreover, $\partial_t u$ cannot vanish on a positive area subset of Q_T without contradicting (1.8).

LEMMA 1.4. *Let $u(x,t)$ solve the direct problem for admissible coefficients and input. Then $h(t) = u(0,t)$ satisfies $h(0) = u_0$ and $h'(t) > 0$ for $0 < t < T$.*

Proof. Applying the reasoning of Lemma 1.3 to the initial boundary value problem

$$\begin{gathered} \partial_t a(u) = \partial_{xx} b(u) \quad \text{for } -1 < x < 1, 0 < t < T, \\ u(x,0) = u_0 \quad \text{for } -1 < x < 1, \\ u(-1,t) = u(1,t) = f(t) \quad \text{for } 0 < t < T \end{gathered}$$

shows that $\partial_t u(x,t) > 0$ in $(-1,1) \times (0,T)$. In particular, $\partial_t u(0,t) = h'(t) > 0$, $0 < t < T$.

Lemma 1.4 supplies necessary conditions for a function h to lie in the range of the coefficient-to-data mapping $(C,K) \mapsto u(0,t)$.

2. Integral identities. Several integral identities relating the unknown coefficients in (1.1) to the overspecified quantities $g(t)$ and $h(t)$ in (1.2) are derived in this section. The first identities relate changes in the *conductivity coefficient*, $K(u)$, to changes in the overspecified data.

THEOREM 2.1. *Let $u = u(x,t)$ be a smooth solution of the direct problem* (1.1) *for admissible coefficients $a(s) = s$ and $b(s) = b_1(s)$ and admissible input f. Suppose $v = v(x,t)$ solves the same problem for different admissible coefficients $a(s) = s$ and $b(s) = b_2(s)$. Then for any $\tau, 0 < \tau \leq T$,*

$$\int_0^\tau (h_1(t) - h_2(t))\vartheta(t)\,dt = \int\int_{Q_\tau} (K_1(v) - K_2(v))\partial_x v \partial_x \psi \,dx\,dt, \tag{2.1}$$

$$\int_0^\tau (g_1(t) - g_2(t))\rho(t)\,dt = \int\int_{Q_\tau} (K_1(v) - K_2(v))\partial_x v \partial_x \varphi \,dx\,dt, \tag{2.2}$$

where

$$p(x,t) = \int_0^1 K_1(v(x,t) + s(u(x,t) - v(x,t)))\,ds \tag{2.3}$$

and where $\psi(x,t)$ solves adjoint problem 3:

$$\begin{gathered} \partial_t \psi + p(x,t)\partial_{xx}\psi = 0 \quad \text{in } Q_\tau, \\ \psi(x,\tau) = 0, \\ p(0,t)\partial_x \psi(0,t) = \vartheta(t) \quad \text{and} \quad \psi(1,t) = 0, \end{gathered} \tag{2.4}$$

and $\varphi(x,t)$ solves the adjoint problem 4:

$$\begin{gathered} \partial_t \varphi + p(x,t)\partial_{xx}\varphi = 0 \quad \text{in } Q_\tau, \\ \varphi(x,\tau) = 0, \\ \partial_x \varphi(0,t) = 0 \quad \text{and} \quad \varphi(1,t) = \rho(t). \end{gathered} \tag{2.5}$$

Proof. Suppose that $u = u(x,t)$ is a smooth solution of the direct problem (1.1) for $a(s) = s$ and $b(s) = b_1(s)$. Suppose also that $v = v(x,t)$ solves the same problem but with $a(s) = s$ and $b(s) = b_2(s)$. Then $w(x,t) = u(x,t) - v(x,t)$ satisfies

$$\begin{gathered} \partial_t w - \partial_{xx}(b_1(u) - b_1(v)) = \partial_{xx}(b_1(v) - b_2(v)) \quad \text{in } Q_T, \\ w(x,0) = 0, \\ \partial_x w(0,t) = 0 \quad \text{and} \quad w(1,t) = 0. \end{gathered} \tag{2.6}$$

Then for any smooth function $\psi(x,t)$ and any τ, $0 < \tau \le T$,

$$\int\int_{Q_\tau} [\partial_t w - \partial_{xx}(b_1(u) - b_1(v))]\psi \, dx \, dt = \int\int_{Q_\tau} \partial_{xx}(b_1(v) - b_2(v))\psi \, dx \, dt.$$

Writing

$$b_1(u) - b_1(v) = (u - v)\int_0^1 b_1'(v + s(u - v)) \, ds = p(u - v),$$

it follows that

$$\begin{aligned} &\int\int_{Q_\tau} w[\partial_t \psi + p(x,t)\partial_{xx}\psi] \, dx \, dt - \int_0^1 w\psi\big|_{t=0}^{t=\tau} dx \\ &\quad + \int_0^\tau (\partial_x(b_1(u) - b_1(v))\psi - (b_1(u) - b_1(v))\partial_x\psi)\big|_{x=0}^{x=1} dt \\ &\quad = \int\int_{Q_\tau} \Delta K(v)\partial_x v \partial_x \psi \, dx \, dt - \int_0^\tau \Delta K(v)\partial_x v \psi\big|_{x=0}^{x=1} dt, \end{aligned} \tag{2.7}$$

where $p(x,t)$ is given by (2.3). The hypotheses on the admissible coefficient K_1 are sufficient to imply that $p(x,t)$ is strictly positive on Q_τ and is Lipschitz with respect to x and t. If $\psi(x,t)$ solves adjoint problem 3, then

$$w\psi\big|_{t=0}^{t=\tau} = 0 \quad \text{and} \quad \partial_x(b_1(u) - b_1(v))\psi = \Delta K(v)\partial_x v \psi = 0 \text{ at } x = 1.$$

Also the boundary conditions on the direct problems for u and v imply that $b_1(u(1,t)) - b_1(v(1,t)) = 0$; $\partial_x b_1(u(0,t)) = \partial_x b_1(v(0,t)) = 0$; and $\Delta K(v)\partial_x v(0,t) = 0$. Then (2.7) reduces to (2.1). On the other hand, if ψ solves the adjoint problem 4, then (2.7) reduces to

$$\begin{aligned} &\int_0^\tau (\partial_x(b_1(u) - b_1(v))\psi(1,t) \, dt \\ &\quad = \int\int_{Q_\tau} \Delta K(v)\partial_x v \partial_x \psi \, dx \, dt - \int_0^\tau \partial_x(b_1(v) - b_2(v))\psi(1,t) \, dt, \end{aligned}$$

i.e.,

$$\int\int_{Q_\tau} \Delta K(v)\partial_x v \partial_x \psi \, dx \, dt = \int_0^\tau \partial_x(b_1(v) - b_2(v))\psi(1,t) \, dt.$$

Since $\partial_x(b_1(v) - b_2(v))\psi(1,t) = (K_1(f) - K_2(f))\partial_x v(1,t)\rho(t) = (g_1(t) - g_2(t))\rho(t)$, we obtain (2.2).

Integral identities relating changes in the *capacity coefficient*, $C(u)$, to changes in the overspecified data can be derived in a similar fashion.

THEOREM 2.2. *Let $u = u(x,t)$ be a smooth solution of the direct problem* (1.1) *for admissible coefficients $a(s) = a_1(s)$ and $b(s) = s$ and admissible input $f(t)$. Suppose $v = v(x,t)$ solves the same problem for different admissible coefficients $a(s) = a_2(s)$ and $b(s) = s$. Then*

$$-\int\int_{Q_\tau} \Delta a(v)\partial_t\psi \, dx\, dt = \int\int_{Q_\tau} \Delta C(v)\partial_t v\psi \, dx\, dt = \int_0^\tau \Delta h(t)\omega(t)\, dt, \tag{2.8}$$

$$-\int\int_{Q_\tau} \Delta a(v)\partial_t\varphi \, dx\, dt = \int\int_{Q_\tau} \Delta C(v)\partial_t v\varphi \, dx\, dt = \int_0^\tau \Delta g(t)\rho(t)\, dt \tag{2.9}$$

where $\psi(x,t)$ solves adjoint problem 5:

$$\begin{gathered} q(x,t)\partial_t\psi + \partial_{xx}\psi = 0 \quad \text{in } Q_\tau, \\ \psi(x,\tau) = 0, \\ \partial_x\psi(0,t) = \omega(t) \quad \text{and} \quad \psi(1,t) = 0, \end{gathered} \tag{2.10}$$

with

$$q(x,t) = \int_0^1 C_1(v(x,t) + s(u(x,t) - v(x,t)))\, ds \tag{2.11}$$

where $\varphi(x,t)$ solves adjoint problem 6:

$$\begin{gathered} q(x,t)\partial_t\varphi + \partial_{xx}\varphi = 0 \quad \text{in } Q_\tau, \\ \varphi(x,\tau) = 0, \\ \partial_x\varphi(0,t) = 0 \quad \text{and} \quad \varphi(1,t) = \rho(t). \end{gathered} \tag{2.12}$$

Proof. Suppose that $u = u(x,t)$ is a smooth solution of the direct problem (1.1) for $a(s) = a_1(s)$ and $b(s) = s$. Suppose also $v = v(x,t)$ solves the same problem but with $a(s) = a_2(s)$ and $b(s) = s$. Then $w(x,t) = u(x,t) - v(x,t)$ satisfies

$$\begin{gathered} \partial_t(a_1(u) - a_1(v)) - \partial_{xx}w = -\partial_t(a_1(v) - a_2(v)) \quad \text{in } Q_T, \\ w(x,0) = 0, \\ \partial_x w(0,t) = 0 \quad \text{and} \quad w(1,t) = 0, \end{gathered} \tag{2.13}$$

and it follows that for any smooth function $\psi(x,t)$,

$$\int\int_{Q_\tau} [\partial_t(a_1(u) - a_1(v)) - \partial_{xx}w]\psi \, dx\, dt = -\int\int_{Q_\tau} \partial_t(a_1(v) - a_2(v))\psi \, dx\, dt.$$

Integrating by parts leads to the identity

$$\begin{aligned} -\int\int_{Q_\tau} w(q(x,t)\partial_t\psi + \partial_{xx}\psi)\, dx\, dt &= \int\int_{Q_\tau} (a_1(v) - a_2(v))\partial_t\psi \, dx\, dt \\ &\quad - \int_0^\tau (\partial_x w\psi - w\partial_x\psi)\big|_{x=0}^{x=1} dt - \int_0^1 (a_1(u) - a_2(v))\psi\big|_{t=0}^{t=\tau} dx; \end{aligned} \tag{2.14}$$

here $q(x,t)$ is given by (2.11). The hypotheses on the admissible coefficient C_1 imply that $q(x,t)$ is strictly positive on Q_τ and q is Lipschitz with respect to x and t. If $\psi(x,t)$ solves adjoint problem 5, then

$$(\psi\partial_x w - w\partial_x\psi)\big|_{x=0}^{x=1} = w(0,t)\partial_x\psi(0,t) \quad \text{and} \quad (a_1(u)-a_2(v))\psi\big|_{t=0}^{t=\tau} = 0$$

and (2.14) becomes (2.8). On the other hand, if ψ solves adjoint problem 6, then

$$(\psi\partial_x w - w\partial_x\psi)\big|_{x=0}^{x=1} = \partial_x w(1,t)\psi(1,t) = (g_1(t)-g_2(t))\rho(t)$$

so that (2.14) reduces to (2.9).

Finally a pair of integral identities can be derived relating simultaneous changes in C and K to changes in the overspecified data.

THEOREM 2.3. *Let $u(x,t;C,K)$ denote the unique solution of the direct problem (1.1) corresponding to admissible coefficients C and K and admissible input $f(t)$. Then*

$$\int_0^\tau (g_1(t)-g_2(t))\vartheta_1(t)\,dt = \int\int_{Q_\tau} \Delta K(v)\partial_x v\partial_x\varphi_1 - \Delta a(v)\partial_t\varphi_1, \tag{2.15}$$

$$\int_0^\tau (h_1(t)-h_2(t))\vartheta_2(t)\,dt = \int\int_{Q_\tau} \Delta K(v)\partial_x v\partial_x\varphi_2 - \Delta a(v)\partial_t\varphi_2, \tag{2.16}$$

where $\varphi_1(x,t)$ solves the adjoint problem 7:

$$\begin{gathered} q(x,t)\partial_t\varphi_1 + p(x,t)\partial_{xx}\varphi_1 = 0 \quad \text{in } Q_\tau, \\ \varphi_1(x,\tau) = 0, \\ \partial_x\varphi_1(0,t) = 0 \quad \text{and} \quad \varphi_1(1,t) = \vartheta_1(t) \end{gathered} \tag{2.17}$$

and $\varphi_2(x,t)$ solves adjoint problem 8:

$$\begin{gathered} q(x,t)\partial_t\varphi_2 + p(x,t)\partial_{xx}\varphi_2 = 0 \quad \text{in } Q_\tau, \\ \varphi_2(x,\tau) = 0, \\ p(0,t)\partial_x\varphi_2(0,t) = \vartheta_2(t) \quad \text{and} \quad \varphi_2(1,t) = 0 \end{gathered} \tag{2.18}$$

for $p(x,t)$ and $q(x,t)$ given by (2.3) *and* (2.11), *respectively.*

Proof. Suppose that $u = u(x,t)$ is a smooth solution of problem (1.1) for $a(s) = a_1(s)$ and $b(s) = b_1(s)$. Suppose $v = v(x,t)$ solves the same problem but with $a(s) = a_2(s)$ and $b(s) = b_2(s)$. Then $w(x,t) = u(x,t) - v(x,t)$ satisfies

(2.19)

$$\begin{gathered} \partial_t(a_1(u)-a_1(v)) - \partial_{xx}(b_1(u)-b_1(v)) = \partial_{xx}(b_1(v)-b_2(v)) - \partial_t(a_1(v)-a_2(v)) \quad \text{in } Q_T, \\ w(x,0) = 0, \\ \partial_x w(0,t) = 0 \quad \text{and} \quad w(1,t) = 0. \end{gathered}$$

Multiplying both sides of the partial differential equation by an arbitrary test function $\varphi(x,t)$ and integrating by parts leads to

(2.20)

$$\begin{aligned} &\int_0^1 (a_1(u)-a_1(v))\varphi\big|_{t=0}^{t=\tau}\,dx - \int\int_{Q_\tau} [(a_1(u)-a_1(v))\partial_t\varphi + (b_1(u)-b_1(v))\partial_{xx}\varphi]\,dx\,dt \\ &\quad - \int_0^\tau [\varphi\partial_x(b_1(u)-b_1(v)) - \partial_x\varphi(b_1(u)-b_1(v))]_{x=0}^{x=1}\,dt \\ &= -\int\int_{Q_\tau} \partial_x(b_1(v)-b_2(v))\partial_x\varphi\,dx\,dt + \int_0^\tau \varphi\partial_x(b_1(v)-b_2(v))\big|_{x=0}^{x=1}\,dt \\ &\quad + \int\int_{Q_\tau} (a_1(v)-a_2(v))\partial_t\varphi\,dx\,dt - \int_0^1 (a_1(v)-a_2(v))\varphi\big|_{t=0}^{t=\tau}\,dx. \end{aligned}$$

Note that

$$(a_1(u) - a_1(v))\partial_t\varphi + (b_1(u) - b_1(v))\partial_{xx}\varphi = w(q(x,t)\partial_t\varphi + p(x,t)\partial_{xx}\varphi),$$

where $p(x,t)$ and $q(x,t)$ are given by (2.3) and (2.11). If the test function $\varphi = \varphi_1$ solves adjoint problem 7, then (2.20) reduces to (2.15). Here use has been made of the facts

$$\partial_x b_1(u) - \partial_x b_2(v) = 0 \text{ at } x = 0 \quad \text{since } \partial_x u(0,t) = \partial_x v(0,t) = 0,$$
$$b_1(u) - b_1(v) = 0 \text{ at } x = 1 \quad \text{since } u(1,t) = v(1,t) = f(t),$$
$$\partial_x(b_1(v) - b_2(v))(1,t) = g_1(t) - g_2(t) \quad \text{and} \quad \partial_x \Delta b(v) = \Delta K(v)\partial_x v.$$

On the other hand, if the test function φ in (2.20) is chosen to equal φ_2, where φ_2 solves adjoint problem 8, then (2.20) reduces to (2.16).

These integral identities can be used to consider the invertibility of the coefficient-to-data mappings in several unknown coefficient inverse problems.

3. Invertibility of the coefficient-to-data mappings. A partial ordering on the metric space $\mathbb{C}[a,b]$ is defined by letting $f < g$ mean that one or the other of the following alternatives holds:

(i) $f(x) < g(x)$ for all x in (a,b), or

(ii) $f(x) = g(x)$ for $a \le x \le p < b$ and $f(x) < g(x)$ for $p < x < b$.

THEOREM 3.1. *Let $u(x,t;K)$ denote the unique solution of the direct initial boundary value problem*

$$\begin{aligned} \partial_t u &= \partial_{xx} b(u) \quad \text{for } 0 < x < 1, 0 < t < T, \\ u(x,0) &= u_0 \quad \text{for } 0 < x < 1, \\ \partial_x u(0,t) = 0 \text{ and } u(1,t) &= f(t) \quad \text{for } 0 < t < T \end{aligned} \tag{3.1}$$

corresponding to admissible coefficient $K(s) = b'(s)$ and admissible input $f(t)$. Let g, h denote the data functions $g(t) = K(f)\partial_x u(1,t;K)$, $h(t) = u(0,t;K)$. Then

(a) *$K_1 > K_2$ on $[u_0, f(T)]$ implies the existence of $\tau, 0 < \tau \le T$, such that $h_1 > h_2$ on $[0,\tau]$.*

(b) *$h_1 > h_2$ on $[0,T]$ implies the existence of $\mu, u_0 < \mu \le f(T)$, such that $K_1 > K_2$ on $[u_0,\mu]$.*

(c) *$K_1 > K_2$ on $[u_0, f(T)]$ implies the existence of $\tau, 0 < \tau \le T$, such that $g_1 > g_2$ on $[0,\tau]$.*

(d) *$g_1 > g_2$ on $[0,T]$ implies the existence of $\mu, u_0 < \mu \le f(T)$, such that $K_1 > K_2$ on $[u_0,\mu]$.*

Proof. Let $u(x,t)$, $v(x,t)$ denote solutions of (3.1) corresponding to admissible conductivity coefficients K_1 and K_2, respectively. Then for $h_1(t) = u(0,t)$ and $h_2(t) = v(0,t)$, it follows from Theorem 2.1 that for any $0 < \tau \le T$,

$$\int_0^\tau (h_1(t) - h_2(t))\vartheta(t)\,dt = \int\int_{Q_\tau} (K_1(v) - K_2(v))\partial_x v \partial_x \psi \,dx\,dt, \tag{3.2}$$

where $\psi(x,t)$ solves adjoint problem 3 and the function $p(x,t)$, given by (2.3), is strictly positive on Q_τ. If the data $\vartheta(t)$ from (2.4), adjoint problem 3, is chosen such that $\vartheta(\tau) = 0$ and $\vartheta(t) > 0$ for $0 < t < \tau$, then one can show that $\psi(x,t)$ must satisfy $\partial_x \psi(x,t) > 0$ on the open region Q_τ. The hypotheses on $f(t)$ imply via Lemma 1.2 that $\partial_x v(x,t) > 0$ almost everywhere in Q_τ.

To prove (a) suppose that $K_1(v) > K_2(v)$ in $\mathbb{C}[u_0, f(T)]$. In particular, suppose $K_1(v) > K_2(v)$ for all v, $u_0 < v < f(T)$. Then $\Delta K(v(x,t)) > 0$ for (x,t) in Q_τ for all τ, $\tau \leq T$, and the right side of (3.2) is positive for any $\tau \leq T$. If there were a point t_0 in $(0,\tau)$ where $\Delta h(t) = h_1(t) - h_2(t)$ were negative, then Δh would be negative in an open neighborhood of t_0. But then we could choose the values of $\vartheta(t)$ on this neighborhood to be so large and positive that the left side of (3.2) would be negative, contradicting equation (3.2). Similarly, if $\Delta h(t) = 0$ for all t in $(0,\tau)$ the integral on the left in (3.2) would be zero. But the right side of (3.2) is positive, and it follows that $\Delta h(t) \geq 0$ on $(0,\tau)$ and $\Delta h(t) > 0$ on a subinterval of positive length in $(0,\tau)$. By decreasing τ if necessary one can ensure that $\Delta h(t) > 0$ for $0 < t < \tau$ or else $\Delta h(t) = 0$ for $0 \leq t \leq \tau' < \tau$ and $\Delta h(t) > 0$ for $\tau' < t < \tau$. If $K_1(v) = K_2(v)$ for $u_0 \leq v \leq u_1$ and $K_1(v) > K_2(v)$ for $u_1 < v < f(T)$, then (3.2) implies $\Delta h(t) = 0$ for $0 \leq t \leq t_1$ where $f(t_1) = u_1$. Now one proceeds as before to establish the existence of a $\tau \leq T$ such that $h_1 > h_2$ on $[0,\tau]$. This proves (a).

Similarly, if $g_1(t) = K_1(f(t))\partial_x u(1,t)$ and $g_2(t) = K_2(f)\partial_x v(1,t)$, then Theorem 2.1 implies that for any $\tau, 0 < \tau \leq T$,

$$\int_0^\tau (g_1(t) - g_2(t))\rho(t)\,dt = \int\int_{Q_\tau} (K_1(v) - K_2(v))\partial_x v \partial_x \varphi \,dx\,dt \tag{3.3}$$

where $\varphi(x,t)$ solves (2.5), adjoint problem 4. If the data $\rho(t)$ satisfies $\rho(\tau) = 0$ and $\rho'(t) < 0$ for $0 < t < \tau$, then one can show that $\varphi(x,t)$ must satisfy $\partial_x \varphi(x,t) > 0$ on the open region Q_τ. One can proceed as in the proof of (a) to show that $K_1 > K_2$ in $\mathbb{C}[u_0, f(T)]$ implies the existence of a $\tau, 0 < \tau \leq T$, such that $g_1 > g_2$ in $\mathbb{C}[0,\tau]$. This proves (c).

To prove (b) suppose that $h_1 > h_2$ on $[0,T]$. In particular, suppose $h_1(t) > h_2(t)$ for $0 < t < T$. Then, under the previous assumptions on $\vartheta(t)$, the left side of (3.2) is positive for any $\tau, 0 < \tau \leq T$. If there were a $u_1 > u_0$ such that $K_1(v) < K_2(v)$ for $u_0 < v < u_1$, then choosing τ such that $f(\tau) = u_1$ would lead to a negative integral on the right side of (3.2) equal to a positive integral on the left. This contradiction precludes the possibility $K_1(v) < K_2(v)$ for $u_0 < v < u_1$. If one supposes that $K_1(v) = K_2(v)$ for $u_0 \leq v \leq f(T)$, then the right side of (3.2) is zero while the left side is strictly positive, which is another contradiction. Similarly, $K_1(v) = K_2(v)$ for $u_0 \leq v \leq u_1 < f(T)$ and $K_1(v) < K_2(v)$ for $u_1 < v < u_2 \leq f(T)$ leads to a contradiction with (3.2). It follows that there exists some μ, $\mu_0 < \mu \leq f(T)$, such that $K_1(v) > K_2(v)$ for $u_0 < v < \mu$ or else $K_1(v) = K_2(v)$ for $u_0 \leq v \leq u_1 < \mu$ and $K_1(v) > K_2(v)$ for $u_1 < v < \mu$. If $h_1(t) = h_2(t)$ for $0 \leq t \leq t_1 < T$ and $h_1(t) > h_2(t)$ for $t_1 < t < T$, then (3.2) can be used to conclude that $K_1(v) = K_2(v)$ for $\mu_0 \leq v \leq u_1 = f(t_1)$. Now we proceed as before to establish the existence of a $\mu > u_1$ such that $K_1 > K_2$ on $[u_0, \mu]$. This proves (b). The proof of (d) is based on (3.3) instead of (3.2) but is otherwise similar.

The properties (a) through (d) are sometimes described by saying that the coefficient-to-data mappings $K \mapsto g$ and $K \mapsto h$ are *monotone* mappings. Note that if the ordering relation $f < g$ were a total ordering on $\mathbb{C}[0,T]$ and $\mathbb{C}[u_0, f(T)]$, then the monotonicity of the coefficient-to-data mappings would imply unicity for solutions of the inverse problem and it would imply existence of an inverse for the coefficient-to-data mapping. Since we have only a partial ordering, the monotonicity implies a weaker version of uniqueness and invertibility.

For f_1, f_2 in $\mathbb{C}[a,b]$, the functions f_1, f_2 are said to be *distinguishable* on $[a,b]$ when there exists a partition $\Pi_n = \{a = \xi_0 < \xi_1 < \cdots < \xi_n = b\}$ of $[a,b]$ such that for each subinterval (ξ_{i-1}, ξ_i), $1 \leq i \leq n$, one has $f_1 < f_2$ or else $f_2 < f_1$ on

(ξ_{i-1}, ξ_i). Two functions are not distinguishable on $[a,b]$ if they are identical on $[a,b]$ or if their graphs cross infinitely often in $[a,b]$. Of course continuous functions that are not distinguishable on $[a,b]$ are not necessarily identical on $[a,b]$. If $D[a,b]$ denotes the subspace of $\mathbb{C}[a,b]$ composed of distinguishable functions, then $D[a,b]$ contains the functions that are analytic on $[a,b]$ but excludes rapidly oscillating functions like $\sin(1/x)$ on $(0,1)$.

Note that f_1 distinguishable from f_2 on $[a,b]$ does not necessarily imply that either of the relations $f_1 > f_2$ or $f_2 > f_1$ holds across the whole interval $[a,b]$. However, exactly one of these relations must hold on (ξ_0, ξ_1), the first subinterval of a partition associated with the distinguishable functions f_1, f_2. If K_1, K_2 are admissible conductivity coefficients that are distinguishable on $[u_0, f(T)]$, then their graphs are ordered on (ξ_0, ξ_1).

COROLLARY 3.2. *Under the hypotheses of Theorem* 3.1, *if* K_1 *is distinguishable from* K_2 *in* $\mathbb{C}[u_0, f(T)]$, *then there exists* $\tau, 0 < \tau \leq T$, *such that* $h_1(t)$ *is distinguishable from* $h_2(t)$ *in* $\mathbb{C}[0,\tau]$ *and* $g_1(t)$ *is distinguishable from* $g_2(t)$ *in* $\mathbb{C}[0,\tau]$.

Proof. Let $\{u_0 < u_1 < \cdots < u_n = f(T)\}$ denote a partition associated with K_1, K_2 in $D[u_0, f(T)]$. Suppose, for example, $K_1 > K_2$ on $u_0 < v < u_1 = f(t_1) \leq f(T)$. Then part (c) of the theorem implies the existence of a $\tau \leq T$ such that $g_1 > g_2$ in $\mathbb{C}[0,\tau]$; i.e., g_1 and g_2 are distinguishable in $\mathbb{C}[0,\tau]$. Similarly, part (a) of the theorem implies h_1 and h_2 are distinguishable on $[0,\tau]$ for some τ, $0 < \tau \leq T$. This is a uniqueness result asserting that coefficients that are distinguishable cannot produce indistinguishable data; i.e., if the inverse problems where $K(u)$ is to be identified from either $h(t)$ or from $g(t)$ has more than one solution, these solutions are not distinguishable on $[u_0, f(T)]$.

The monotonicity of the coefficient-to-data mappings also implies a restricted type of invertibility. One way of using the integral identities to implement this inversion is illustrated in §4.

THEOREM 3.3. *Let* $u(x,t;C)$ *be the unique solution of the direct initial boundary value problem*

$$\begin{aligned} \partial_t a(u) &= \partial_{xx} u \quad \textit{for } 0 < x < 1, 0 < t < T, \\ u(x,0) &= u_0 \quad \textit{for } 0 < x < 1, \\ \partial_x u(0,t) &= 0 \textit{ and } u(1,t) = f(t) \quad \textit{for } 0 < t < T \end{aligned} \tag{3.4}$$

corresponding to admissible coefficient $C(s) = a'(s)$ *and admissible input* $f(t)$. *Then in the notation of Theorem* 3.1,

(a) $C_1 > C_2$ *on* $[u_0, f(T)]$ *implies the existence of* τ, $0 < \tau \leq T$, *such that* $h_2 > h_1$ *on* $[0,\tau]$.

(b) $h_1 > h_2$ *on* $[0,T]$ *implies the existence of* μ, $u_0 < \mu \leq f(T)$, *such that* $C_2 > C_1$ *on* $[u_0, \mu]$.

(c) $C_1 > C_2$ *on* $[u_0, f(T)]$ *implies the existence of* $\tau, 0 < \tau \leq T$, *such that* $g_1 > g_2$ *on* $[0,\tau]$.

(d) $g_1 > g_2$ *on* $[0,T]$ *implies the existence of* μ, $u_0 < \mu \leq f(T)$, *such that* $C_1 > C_2$ *on* $[u_0, \mu]$.

Proof. The proof of this theorem is based on the identities (2.8) and (2.9) instead of (2.1) and (2.2) but otherwise is quite similar to the proof of Theorem 3.1. If the data $\omega(t)$ in (2.10) is chosen to be positive for $0 < t < \tau$, then the solution ψ of adjoint problem 5 is necessarily negative in Q_τ. Using these facts in (2.8) leads to results (a) and (b) that $C_1 > C_2$ is consistent with $h_2 > h_1$.

There appears to be no natural way to extend the notion of monotonicity to the coefficient-to-data mapping associated with the simultaneous identification of the two unknown coefficients $C(u)$ and $K(u)$ (or, alternatively, $a(u)$ and $b(u)$) from the pair of overspecifications (1.2). However, restricted uniqueness and invertibility results are true for this problem as they were in the inverse problems in a single unknown. The restricted inversion for this problem will be discussed in the next section.

THEOREM 3.4. *Let $u(x,t)$, $v(x,t)$ denote solutions of the direct problem* (1.1) *corresponding to admissible coefficient pairs (C_1,K_1), (C_2,K_2), respectively, with admissible input $f(t)$. Let (g_1,h_1) and (g_2,h_2) denote the corresponding data pairs as defined in* (1.2).

(a) *If K_1 is distinguishable from K_2 and C_1 is distinguishable from C_2 in $\mathbb{C}[u_0,f(T)]$, then the coefficient pairs generate data pairs (g_1,h_1) and (g_2,h_2) that are distinguishable on $[0,t_1]$ for some t_1, $0<t_1\le T$.*

(b) *If the data pairs (g_1,h_1) and (g_2,h_2) are identical on $[0,T]$, then K_1 is not distinguishable from K_2 and C_1 is not distinguishable from C_2 on $[u_0,f(T)]$.*

Proof. Suppose that K_1 is distinguishable from K_2 and C_1 is distinguishable from C_2 in $\mathbb{C}[u_0,f(T)]$, and let $\{u_0<u_1<\cdots<u_n=f(T)\}$ denote a partition associated with K_1, K_2 in $D[u_0,f(T)]$ and $\{u_0<u_i'<\cdots<u_n=f(T)\}$ denote a partition associated with C_1, C_2. Choose the smaller of the two numbers u_i, $_1'$ and call this u_1.

There are several cases to consider but the idea of the proof can be seen by considering just two of them. Suppose, for example, that $K_1(v)>K_2(v)$ and $a_1(v)<a_2(v)$ for $u_0<v<u_1=f(t_1)\le f(T)$. If $\vartheta_1(t)$ in adjoint problem 7 satisfies $\vartheta_1(\tau)=0$ and $\vartheta_1(t)>0$ for $0<t<\tau$, then we can show that $\partial_x\varphi_1(x,t)$ and $\partial_t\varphi_1(x,t)$ are both positive in Q_τ. Similarly if $\vartheta_2(t)$ in adjoint problem 8 satisfies $\vartheta_2(\tau)=0$ and $\vartheta_2(t)>0$ for $0<t<\tau$, then we can show $\partial_x\varphi_2(x,t)>0$ and $\partial_t\varphi_2(x,t)<0$ in Q_τ. If we choose $\tau=t_1$, then $\Delta K(v)\partial_x v\partial_x\varphi_1-\Delta a(v)\partial_t\varphi_1>0$ on Q_τ and it follows by the arguments used in Theorem 3.1, that (2.15) implies $g_1>g_2$ on $(0,t_1)$. On the other hand, if we have $K_1(v)>K_2(v)$ and $a_1(v)>a_2(v)$ for $u_0>v<u_1=f(t_1)\le f(T)$, then $\Delta K(v)\partial_x v\partial_x\varphi_2-\Delta a(v)\partial_t\varphi_2>0$ on Q_τ. In this case the usual arguments used with (2.16) imply $h_1>h_2$ on $(0,t_1)$. We have proved that coefficient pairs $(C_1,K_1),(C_2,K_2)$ that are distinguishable cannot produce data pairs (g,h) that are not distinguishable; this implies uniqueness of the solution to the inverse problem for the simultaneous identification of (C,K) (equivalently, (a) and (b) from the data in (1.2).

To prove (b), suppose K_1 is distinguishable from K_2 and C_1 is distinguishable from C_2 in $\mathbb{C}[u_0,f(T)]$, and suppose also that $\Delta g(t)=\Delta h(t)=0$ on $[0,T]$. Then $\Delta g(t)=\Delta h(t)=0$ on $(0,t_1)$ where $u_1=f(t_1)$ where u_1 is as it was in the proof of part (a) so that $\Delta a(v)$ and $\Delta K(v)$ do not change sign on (u_0,u_1). Then (2.15) implies

$$\int\int_{Q_\tau}\Delta K(v)\partial_x v\partial_x\varphi_1\,dx\,dt=\int\int_{Q_\tau}\Delta a(v)\partial_t\varphi_1\,dx\,dt.$$

Since $\partial_x v\partial_x\varphi_1$ and $\partial_t\varphi_1$ are positive on Q_τ and $\Delta K(v)$ and $\Delta a(v)$ do not change sign on Q_τ, $\Delta K(v)$ must have the same sign as $\Delta a(v)$ on (u_0,u_1). In the same way (2.16) implies

$$\int\int_{Q_\tau}\Delta K(v)\partial_x v\partial_x\varphi_2\,dx\,dt=\int\int_{Q_\tau}\Delta a(v)\partial_t\varphi_2\,dx\,dt,$$

and in view of the fact that $\partial_t\varphi_2\le 0$ while $\partial_x v\partial_x\varphi_2\ge 0$, the same argument now shows that $\Delta K(v)$ and $\Delta a(v)$ have opposite signs on (u_0,u_1). It follows that ΔK and Δa are both zero on (u_0,u_1) and, since $\Delta g=\Delta h=0$ on $(0,T)$, the interval (u_0,u_1) must cover the whole interval $(u_0,f(T))$. Since the assumption that the coefficient pairs

are distinguishable leads to the contradictory conclusion that the coefficient pairs are identical on $(u_0, f(T))$, it follows that the coefficient pairs must be indistinguishable on $(u_0, f(T))$. Statement (b) asserts that the coefficient-to-data mapping $(C, K) \mapsto (g, h)$ is injective, but (a) and (b) are not quite equivalent since indistinguishable is not the same as identical.

4. Invertibility in the space of polygonal approximations. Let $u(x,t; K)$ denote the unique solution of the direct initial boundary value problem (3.1) corresponding to the admissible conductivity coefficient K and admissible input $f(t)$. Then the maximum principle implies $u_0 \leq u(x,t) \leq f(T)$ for all (x,t) in Q_T and we may seek to determine $K(u)$ on the interval $u_0 \leq u \leq f(T)$ from either piece of overspecified data $g(t) = K(f(t))\partial_x u(1,t)$ or $h(t) = u(0,t)$.

The monotonicity of the coefficient-to-data mapping can be exploited in constructing an inverse mapping if the domain and range are totally ordered. The spaces $\mathbb{C}[0,T]$ and $\mathbb{C}[u_0, f(T)]$ are only partially ordered, but a total ordering can be induced by restricting the mappings to subspaces that consist of piecewise linear continuous functions. The admissibility of $f(t)$ is essential for this step.

Let $\{0 = t_0 < t_1 < \cdots < t_n = T\}$ denote a partition of the interval $[0,T]$. Then $\{u_0 < u_1 < \cdots < u_n = f(T)\}$, where $f(t_j) = u_j$ for $j = 0, 1, \ldots, n$, defines a corresponding partition of the interval $[u_0, f(T)]$. Define $P_n K(u)$, a polygonal (i.e., continuous and piecewise linear) approximation to $K(u)$ on $[u_0, u_n]$, as follows. For each m, $m = 1, \ldots, n$, let $P_n K(u)$ be given for $u_{m-1} \leq u \leq u_m$ by

$$P_n K(u) = \kappa_{m-1} \frac{u_m - u}{u_m - u_{m-1}} + \kappa_m \frac{u - u_{m-1}}{u_m - u_{m-1}}. \tag{4.1}$$

The constant parameters $\{\kappa_0, \ldots, \kappa_n\}$ can be determined from either of the data pairs f and g or f and h. For purposes of this illustration f and g are used. Assume the initial constant, $\kappa_0 = K(u_0)$, is known. This represents no loss of generality since κ_0 can be obtained from the data $f(t), g(t)$; see [7], [8].

First it will be shown how κ_1 is obtained and then the procedure will be generalized to the case of subsequent κ's. Let $K_1(u)$ equal the linear function given by (4.1) on (u_0, u_1). Note that the definition of $K_1(u)$ contains the as-yet unknown parameter κ_1. Let $K_2(u)$ be given by (4.2) with $\kappa_1 = \kappa_0$; i.e., $K_2(u)$ equals the constant κ_0 for u in (u_0, u_1). Note that the family of functions $K_i(u)$ that are linear on (u_0, u_1) and satisfy $K_i(u_0) = \kappa_0$ is totally ordered.

Let $g_1(t)$ equal the overprescribed data function $g(t)$ on $(0, t_1)$, and let $g_2(t) = K_2(f(t))\partial_x v(1,t; K_2)$ on $(0, t_1)$, where $v(x,t; K_2)$ denotes the solution of the direct problem (3.1) on $Q_{t_1} = (0,1) \times (0, t_1)$ corresponding to the coefficient $K = K_2$. Note that Corollary 3.2 implies $\Delta g(t) = g_1(t) - g_2(t)$ is not zero unless the linear functions K_1 and K_2 are identical on (u_0, u_1). Finally let $\varphi(x,t)$ denote the solution of the adjoint problem 4 for data $\rho(t)$ such that $\rho(t_1) = 0$ and $\rho'(t) < 0$ for $0 < t < t_1$. Then Theorem 2.1 implies

$$\int\int_{Q_{t_1}} \Delta K(v) \partial_x v \partial_x \varphi \, dx \, dt = \int_0^{t_1} \Delta g(t) \rho(t) \, dt.$$

But

$$\Delta K(v) = K_1(v) - K_2(v) = (\kappa_1 - \kappa_0) \frac{v - u_0}{u_1 - u_0} = (\kappa_1 - \kappa_0) \Lambda_1(v) \quad \text{for } u_0 \leq v \leq u_1,$$

which leads at once to the result

$$\kappa_1 = \kappa_0 + \frac{\int_0^{t_1} \Delta g(t)\rho(t)\,dt}{\int\int_{Q_{t_1}} \Lambda_1(v)\partial_x v \partial_x \varphi\,dx\,dt}. \tag{4.2}$$

Since the integral in the denominator can only be positive, κ_1 is greater than or less than κ_0 according to whether $\Delta g(t)$ is positive or negative.

To find κ_m once the first m constants $\kappa_0, \kappa_1, \ldots, \kappa_{m-1}$ are known, let $K_1(u)$ and $K_2(u)$ both be given by (4.1) for $u_0 \le u \le u_{m-1}$. Then $\Delta K(v(x,t)) = 0$ for (x,t) in $(0,1) \times (0, t_{m-1})$. On (u_{m-1}, u_m) let

$$K_1(u) = \kappa_{m-1}\frac{u_m - u}{u_m - u_{m-1}} + \kappa_m \frac{u - u_{m-1}}{u_m - u_{m-1}},$$
$$K_2(u) = \kappa_{m-1}.$$

Let $v(x,t)$ denote the solution on $(0,1) \times (0,t_m) = Q_{t_m}$ of the direct problem (3.1) for $K(v) = K_2(v)$. Then $\Delta g(t)$ denotes the difference between the data function $g_1(t) = g(t)$ and $g_2(t) = K_2(f(t))\partial_x v(1,t)$ for $0 < t < t_m$. Finally let $\varphi(x,t)$ denote the solution of the adjoint problem (2.5) on Q_{t_m}. Then Theorem 2.1 leads, as it did for $m = 1$, to the result

$$\kappa_m = \kappa_{m-1} + \frac{\int_0^{t_m} \Delta g(t)\rho(t)\,dt}{\int\int_{S_m} \Lambda_m(v)\partial_x v \partial_{x\varphi}\,dx\,dt}, \tag{4.3}$$

where $S_m = (0,1) \times (t_{m-1}, t_m)$ is the part of Q_{t_m} where $\Delta K(v(x,t))$ is not zero. Executing this procedure for $m = 1$ to $m = n$ generates a polygonal approximation $P_n K(u)$ for $K(u)$ on the partition $\{u_0 < u_1 < \cdots < u_n\}$ of $[u_0, f(T)]$.

This construction of the polygonal coefficient $P_n K(u)$ is explicit and can be carried out for every partition of $[u_0, f(T)]$. This reflects the invertibility of the coefficient-to-data mapping $K(u) \mapsto g$ in the space of polygonal coefficients where the relation $f < g$ acts as a total ordering. Numerical experiments to test the effectiveness of algorithms that incorporate this construction will be discussed in future publications.

Now consider the simultaneous determination of the two unknown coefficients C and K. Note that when determining just one of the coefficients, assuming the other to be known, one is free to choose either of the overspecifications in (1.2). However, to determine both of the coefficients simultaneously one must use both of the overspecifications in (1.2).

Let $u = u(x,t;C,K)$ denote the unique solution of (1.1) corresponding to admissible coefficients C and K and admissible input, $f(t)$. Let $h(t), g(t)$ be as defined in (1.2).

With a partition of $[u_0, f(T)]$ as previously chosen, constants κ_m, ϑ_m are to be determined so as to define $P_n K(u)$ and $P_n a(u)$, polygonal (i.e., continuous and piecewise linear) approximations to $K(u)$ and $a(u)$ on $[u_0, u_n]$. For each m, $m = 1, \ldots, n$, $P_n K(u)$ and $P_n a(u)$ are defined for $u_{m-1} \le u \le u_m$ by

$$P_n K(u) = \kappa_{m-1}\frac{u_m - u}{u_m - u_{m-1}} + \kappa_m \frac{u - u_{m-1}}{u_m - u_{m-1}}, \tag{4.4}$$

$$P_n a(u) = \vartheta_{m-1}\frac{u_m - u}{u_m - u_{m-1}} + \vartheta_m \frac{u - u_{m-1}}{u_m - u_{m-1}}. \tag{4.5}$$

One can define an algorithm for determining the constants κ_m and ϑ_m for $m = 1, \ldots, n$, assuming that κ_0 and ϑ_0 are known from the data. The algorithm proceeds step by step for each m. For $m > 0$ fixed, suppose $P_n K(u)$ and $P_n a(u)$ are given by (4.4) and (4.5) with κ_j and ϑ_j known for $j = 1$ up to $m-1$. To extend $P_n K(u)$ and $P_n a(u)$ to (u_{m-1}, u_m) define $K_2(v) = \kappa_{m-1}$ and $a_2(v) = \vartheta_{m-1}$ for $u_{m-1} \le v \le u_m$. Also define

$$K_1(v) = \kappa_{m-1}\frac{u_m - v}{u_m - u_{m-1}} + \kappa_m \frac{v - u_{m-1}}{u_m - u_{m-1}}, \qquad u_{m-1} \le v \le u_m; \tag{4.6}$$

$$a_1(v) = \vartheta_{m-1}\frac{u_m - v}{u_m - u_{m-1}} + \vartheta_m \frac{v - u_{m-1}}{u_m - u_{m-1}}, \qquad u_{m-1} \le v \le u_m, \tag{4.7}$$

noting that the functions $K_1(v)$ and $a_1(v)$ depend on the (as-yet unknown) quantities κ_m and ϑ_m. Let $v(x,t)$ denote the solution on Q_{t_m} of (1.1) for coefficients $K = K_2$, $a = a_2$ and the same input $f(t)$, and define functions

$$g_2(t) = K_2(f(t))\partial_x v(1,t) \quad \text{and} \quad h_2(t) = v(0,t), \qquad 0 \le t \le t_m.$$

Then with $g_1(t), h_1(t)$ equal to the overspecified data $g(t), h(t)$, respectively, (2.15) and (2.16) are used to write

$$\int_0^{t_m} \Delta g(t)\vartheta_1(t)\,dt = \int\int_{S_m} [\Delta K(v)\partial_x v \partial_x \varphi_1 - \Delta a(v)\partial_t \varphi_1]\,dx\,dt, \tag{4.8}$$

$$\int_0^{t_m} \Delta h(t)\vartheta_2(t)\,dt = \int\int_{S_m} [\Delta K(v)\partial_x v \partial_x \varphi_2 - \Delta a(v)\partial_t \varphi_2]\,dx\,dt, \tag{4.9}$$

where $S_m = (0,1)\times(t_{m-1}, t_m)$. Here use was made of the fact that $\Delta K(v) = \Delta a(v) = 0$ for $u_0 \le v \le u_{m-1}$. In addition, $\varphi_1(x,t)$ and $\varphi_2(x,t)$ denote the solutions of adjoint problems 7 and 8, respectively, for coefficients $p(x,t), q(x,t)$ given by (2.3) and (2.11) with $K = K_2$ and $a = a_2$. Then (4.8), (4.9) are two equations in the two unknowns κ_m, ϑ_m. These equations assume the form

$$\begin{bmatrix} \kappa_m \\ \vartheta_m \end{bmatrix} = \begin{bmatrix} \kappa_{m-1} \\ \vartheta_{m-1} \end{bmatrix} + M^{-1} \begin{bmatrix} \Delta G_m \\ \Delta H_m \end{bmatrix}, \tag{4.10}$$

where

$$\Delta G_m = \int_0^{t_m} \Delta g(t)\vartheta_1(t)\,dt, \qquad \Delta H_m = \int_0^{t_m} \Delta h(t)\vartheta_2(t)\,dt,$$

and the entries M_{ij} of the 2-by-2 matrix M are given by

$$M_{11} = \int\int_{S_m} \Lambda_m(v)\partial_x v \partial_x \varphi_1\,dx\,dt, \quad M_{12} = -\int\int_{S_m} \Lambda_m(v)\partial_t \varphi_1\,dx\,dt,$$
$$M_{21} = \int\int_{S_m} \Lambda_m(v)\partial_x v \partial_x \varphi_2\,dx\,dt, \quad M_{22} = -\int\int_{S_m} \Lambda_m(v)\partial_t \varphi_2\,dx\,dt$$

with

$$\Lambda_n(v) = \frac{v - u_{n-1}}{u_n - u_{n-1}} \quad \text{for } u_{n-1} \le v \le u_n.$$

Note that $\Lambda_m(v) \geq 0$ on S_m; hence $M_{11} > 0$, $M_{22} > 0$, $M_{21} > 0$, and $M_{12} < 0$. Then $\det M$ is positive and a unique solution for κ_m, ϑ_m exists independent of the data g, h. This is a reflection of the invertibility of the coefficient to data map $(a, K) \mapsto (g, h)$ within the space of polygonal coefficients.

Note that for test function $\varphi(x, t)$ that solves either of the adjoint problems 7 or 8, it follows that

$$\begin{aligned}\int\int_{Q_\tau} \Delta a(v)\partial_t\varphi \, dx \, dt &= \int_0^1 \Delta a(v)\varphi\big|_{t=0}^{t=\tau} dx - \int\int_{Q_\tau} \partial_t \Delta A(v)\varphi \, dx \, dt \\ &= 0 - \int\int_{Q_\tau} \Delta C(v)\varphi\partial_t v \, dx \, dt.\end{aligned}$$

Then (4.8), (4.9) become

$$\int_0^{t_m} \Delta g(t)\vartheta_1(t) \, dt = \int\int_{S_m} [\Delta K(v)\partial_x v\partial_x\varphi_1 + \Delta C(v)\partial_t v\varphi_1] \, dx \, dt, \tag{4.11}$$

$$\int_0^{t_m} \Delta h(t)\vartheta_2(t) \, dt = \int\int_{S_m} [\Delta K(v)\partial_x v\partial_x\varphi_2 + \Delta C(v)\partial_t v\varphi_2] \, dx \, dt. \tag{4.12}$$

These equations provide the basis for the algorithm to generate polygonal approximations for K and C rather than K and a.

REFERENCES

[1] G. Chavent and P. Lemonnier, *Identification de la nonlinearite d'une equation parabolique quasilineare*, Appl. Math. Optim., 1 (1971), pp. 121–162.

[2] A. J. Silva Neto and N. N. Ozisik, *Estimation of space and time dependent strength of a volumetric heat source in a 1-dimensional plate*, Internat. J. Heat Transf., to appear.

[3] U. Hornung, *Identification of nonlinear soil parameters from an input-output experiment*, in Prog. Sci. Comput., Birkhäuser Boston, Cambridge, MA, 1983.

[4] P. C. DuChateau, *Monotonicity and uniqueness results in identifying an unknown coefficient in a nonlinear diffusion equation*, SIAM J. Appl. Math., 41 (1981), pp. 310–323.

[5] P. C. DuChateau and W. Rundell, *Unicity in an inverse problem for an unknown reaction term in a reaction-diffusion equation*, J. Differential Equations, 59 (1985), pp. 155–164.

[6] N. V. Muzylev, *Uniqueness theorems for some converse problems of heat conduction*, USSR Comput. Math. Math. Phys., 20 (1986), pp. 120–134.

[7] J. R. Cannon and P. C. DuChateau, *Determination of unknown coefficients in parabolic operators from overspecified initial boundary data*, J. Heat Transfer, 100 (1978), pp. 503–507.

[8] J. R. Cannon and P. C. DuChateau, *Some asymptotic boundary behavior of solutions of nonlinear parabolic initial boundary value problem*, J. Math. Anal. Appl., 68 (1979), pp. 536–547.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1488–1513, November 1995

007

ANALYTICITY OF SOLUTIONS OF THE GENERALIZED KORTEWEG–DE VRIES EQUATION WITH RESPECT TO THEIR INITIAL VALUES*

BING-YU ZHANG†

Abstract. The initial value problem (IVP) of the generalized Korteweg–de Vries (KdV) equation

$$\partial_t u + \partial_x(a(u)) + \partial_x^3 u = 0, \qquad u(x,0) = \phi(x)$$

is well posed in the classical Sobolev space $H^s(R)$ with $s > 3/4$, which establishes a nonlinear map K from $H^s(R)$ to $C([-T,T];H^s(R))$. It is shown that

(i) if $a = a(x)$ is a C^∞ function on R to R, then K is infinitely many times Frechet differentiable;

(ii) if $a = a(x)$ is a polynomial, then K is analytic, i.e., for any $\phi \in H^s(R)$, K has a Taylor series expansion

$$K(\phi + h) = \sum_{n=0}^{\infty} \frac{1}{n!} K^{(n)}(\phi)[h^n].$$

Each term $y_n = K^{(n)}(\phi)[h^n]$ in the series solves a linearized KdV equation. Consequently, any "small" perturbation $K(\phi + h)$ of $K(\phi)$ can be obtained by solving a series of linear problems.

The proof of these results relies on various smoothing properties of the associated linear KdV equation.

Key words. well-posedness, Frechet differentiability, analyticity

AMS subject classifications. 35Q20, 35B30, 35C10

1. Introduction.

Consider the initial value problem (IVP) for the generalized Korteweg–de Vries (KdV) equation

$$\begin{cases} \partial_t u + \partial_x(a(u)) + \partial_x^3 u = 0, & t,\ x \in R, \\ u(x,0) = \phi(x), \end{cases} \tag{1.1}$$

in which $a(x)$ is assumed to be a C^∞ function on R to R, although a weaker differentiability suffices for some results below.

The KdV equation (i.e., $a(u) = u^2/2$ in (1.1)) and its generalized form (1.1) have been studied by many authors (see, for example, [1], [2], [13], [17] for an initial collection of references). In particular, it is well known that the IVP (1.1) is locally well posed in the classical Sobolev space $H^s(R)$ with $s > 3/2$ (cf. [12], [13], and [22]) and globally well posed with some restrictions on $a(u)$ or the size of the initial data ϕ when $s \geq 2$ (see [11], [13], and [20] for the well-posedness of the IVP (1.1) in other function spaces). In the case in which $a'(u) = u^k$ in (1.1) with k being a positive integer, Kenig, Ponce, and Vega [18] proved that the IVP (1.1) is locally well posed

* Received by the editors January 1, 1993; accepted for publication (in revised form) March 23, 1994. This research was partially supported by the 1992 Taft Summer Research Grant of the University of Cincinnati.

† Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221.

in the space $H^s(R)$ with

$$\text{(1.2)} \qquad \begin{cases} s > 3/4 & \text{if } k = 1, \\ s \geq 1/4 & \text{if } k = 2, \\ s \geq 1/12 & \text{if } k = 3, \\ s \geq \frac{k-4}{2k} & \text{if } k \geq 4, \end{cases}$$

and globally well posed when $1 \leq k \leq 3$ and $s \geq 1$.[1] Their proof is based on careful analysis of various smoothing properties of the associated linear problem together with the contraction mapping principle.

With a few technical modifications, the approach by Kenig, Ponce, and Vega [15], [18] can be easily adapted to the general case. To state this expected result precisely, we introduce the following Banach spaces as Kenig, Ponce, and Vega did in [18].

Let $s > 0$ and $T > 0$ be given. For

$$w: \quad R \times [-T, T] \to R,$$

define

$$\lambda_1(T, w) = \sup_{[-T,T]} \|w(., t)\|_s,$$

$$\lambda_2(T, w) = \left(\sup_x \int_{-T}^{T} |D^s \partial_x w(x, t)|^2 dt \right)^{1/2},$$

$$\lambda_3(T, w; l) = \left(\int_{-T}^{T} \|J^l \partial_x w(., t)\|_\infty^4 dt \right)^{1/4}$$

with $l \in [0, s - 3/4]$, where $J^s = (1 - \partial_x^2)^{s/2}$,

$$\lambda_4(T, w; r) = (1 + T)^{-\rho} \left(\int_R \sup_{[-T,T]} |J^r w(x, t)|^2 dx \right)^{1/2}$$

with $r \in [0, s - 3/4)$ and $\rho > 3/4$ being a fixed constant, and

$$\text{(1.3)} \qquad \Lambda_{l,r}^s(T; w) = \max \{\lambda_1(T, w),\ \lambda_2(T, w),\ \lambda_3(T, w; l),\ \lambda_4(T, w; r)\}.$$

Define

$$\text{(1.4)} \qquad X_{l,r}^{T,s} = \{w \in C([-T, T]; H^s(R)) \mid \Lambda_{l,r}^s(T; w) < \infty\}$$

[1] In the case of the classical KdV equation, Bourgain [4], [5] has recently shown that the IVP (1.1) is globally well posed in L^2. Then, combining the estimates established in [18] with some ideas introduced by Bourgain, Kenig, Ponce, and Vega [19] proved that the IVP (1.1) is locally well posed in H^{-s} with $s < 5/8$.

with $(l,r) \in [0, s-3/4] \times [0, s-3/4)$. It is a Banach space equipped with the norm

$$\|w\|_{X_{l,r}^{T,s}} := \Lambda_{l,r}^{s}(T;w).$$

It can be shown that if $a(0) = a'(0) = 0$, then for any $\phi \in H^s(R)$, there exists a $T > 0$ depending only on $\|\phi\|_s$ such that the IVP (1.1) has a unique solution $u \in X_{l,r}^{T,s}$, where $s > 3/4$ and $(l,r) \in [0, s-3/4] \times [0, s-3/4)$. The global result can be also obtained when the same restrictions are enforced on $a(u)$ in (1.1) or on the size of the initial value ϕ as those in Kato [13] but with $s \geq 1$ instead of $s \geq 2$.

Thus the IVP (1.1) establishes a nonlinear map K from $H^s(R)$ to $X_{l,r}^{T,s}$ (or $C([-T,T]; H^s(R))$). The regularity of this nonlinear map K is our major concern in this paper.

In the case of the classical KdV equation, Bona and Smith [2] first proved that the corresponding map K is continuous from $H^s(R)$ to $C(0,T;H^s(R))$. In [23] Saut and Temam showed that K is locally Hölder continuous with exponent $1/2$, while considering K as a map from $H^{s+1/2}(R)$ to $L^\infty((-T,T);H^s(R))$ $(s \geq 2)$. The continuity of the map K from $H^s(R)$ to $C(0,T;H^s(R))$ in the general case is established by Kato [13]; it follows from his semigroup approach. These early results did not use smoothing properties of the equation. In the case of $a'(u) = u^k$, as a by-product of the fixed point approach based on smoothing properties of the equation, Kenig, Ponce, and Vega [18] proved that K is Lipschitz continuous from $H^s(R)$ to $X_{l,r}^{T,s}$. Later, Zhang proved in [25] that the map K corresponding to the classical KdV equation is infinitely many times Frechet differentiable from $H^s(R)$ to $X_{l,r}^{T,s}$ and it has a Taylor series expansion at any given $\phi \in H^s(R)$. That is, the map K is analytic from $H^s(R)$ to $X_{l,r}^{T,s}$.

In this paper we shall first show that the nonlinear map K established by the IVP (1.1) is also infinitely many times Frechet differentiable from $H^s(R)$ to $X_{l,r}^{T,s}$. For any $n \geq 1$, its nth derivative $K^{(n)}(\phi)$ at $\phi \in H^s(R)$, an n-linear map from the n-fold product space $(H^s(R))^n$ into $X_{l,r}^{T,s}$, can be constructed by solving a system of inhomogeneous linearized KdV equations. More precisely, for any $n \geq 1$ and $h_k \in H^s(R)$ $(k = 1, 2, \ldots, n)$, let

$$w_{[1,\ldots,n]}^{(n)} := K^{(n)}(\phi)[h_1, \ldots, h_n]. \tag{1.5}$$

Then solve

$$\begin{cases} \partial_t w_{[1]}^{(1)} + \partial_x (a'(u) w_{[1]}^{(1)}) + \partial_x^3 w_{[1]}^{(1)} = 0, \\ \\ w_{[1]}^{(1)}(x,0) = h(x) \end{cases} \tag{1.6}$$

for $n = 1$ and

$$\begin{cases} \partial_t w_{[1,\ldots,n]}^{(n)} + \partial_x (a'(u) w_{[1,\ldots,n]}^{(n)}) + \partial_x^3 w_{[1,\ldots,n]}^{(n)} = -\partial_x (H_n), \\ \\ w_{[1,\ldots,n]}^{(n)}(x,0) = 0 \end{cases} \tag{1.7}$$

for $n \geq 2$, where $u = K(\phi)$ and H_n is a polynomial of $w_{[i_1,\ldots,i_j]}^{(j)}$ with $1 \leq i_1, \ldots, i_j \leq n$ and $1 \leq j \leq n-1$ (see §3 for the structure of H_n).

If we choose $h_1 = h_2 = \cdots = h$ and let

$$y_n = K^{(n)}(\phi)[h^n],$$

which is a homogeneous polynomial function of degree n from $H^s(R)$ to $X_{l,r}^{T,s}$, then we can define the nth Taylor polynomial P_n of K at $\phi \in H^s(R)$,

$$\begin{aligned} P_n(\phi)[h] &:= K(\phi) + \sum_1^n \frac{K^{(n)}(\phi)}{k!}[h^k] \\ &= u + \sum_{k=1}^n \frac{y_k}{k!}, \end{aligned} \tag{1.8}$$

and a formal Taylor series P_∞ of K at $\phi \in H^s(R)$ by letting $n \to \infty$ in (1.8).

While assuming that $a(u)$ is a polynomial, we shall prove that for any $\phi \in H^s(R)$, there exists a $\delta > 0$ such that if $h \in H^s(R)$ with $\|h\|_s \le \delta$, then

$$K(\phi + h) = \sum_{k=0}^\infty \frac{K^{(n)}(\phi)}{n!}[h^n] \tag{1.9}$$

with the series converging uniformly about h with $\|h\|_s \le \delta$ in the space $X_{l,r}^{T,s}$. In other words, the map K is analytic from $H^s(R)$ to $X_{l,r}^{T,s}$.

The proof of the above regularity results for the map K relies on various smoothing properties of the associated linear KdV equation (see [6], [8], [9], [13], [16], [18], and [24] for smoothing properties of dispersive wave equations and their applications).

It is worth pointing out that each term $y_n = K^{(n)}(\phi)[h^n]$ in the series (1.9) is a solution of a linearized KdV equation. Hence, in the case in which $a(u)$ in (1.1) is a polynomial, solutions of the IVP (1.1) can be obtained by solving a series of linear problems. This would provide us with a new approach to the IVP (1.1) (cf. [3]). One can first solve system (1.6)–(1.7) inductively and construct the Taylor polynomial P_n (1.8). Then one shows that the $\{P_n\}_1^\infty$ is a Cauchy sequence and its limit ($n \to \infty$) is the needed solution of the IVP (1.1). One of the advantages of this approach is that the following result related to the global well-posedness of the IVP (1.1) follows almost automatically [3]: *For any $s > 3/4$ and $T > 0$, let $\mathcal{D}_{s,T}$ be the collection of all $\phi \in H^s(R)$ such that the corresponding solution of the* IVP (1.1) *exists on the interval $[-T, T]$. Then $\mathcal{D}_{s,T}$ is a nonempty open subset of $H^s(R)$.*

The paper is organized as follows. In §2, in addition to listing the estimates concerning the IVP

$$\partial_t u + \partial_x^3 u = f(x,t), \qquad u(x,0) = u_0(x), \tag{1.10}$$

which are needed to establish the nonlinear results, we consider the IVP for the following linear equation:

$$\begin{cases} \partial_t u + \partial_x(a(v)u) + \partial_x^3 u = f(x,t), \quad x,\ t \in R, \\ u(x,0) = \phi(x). \end{cases} \tag{1.11}$$

We show that if $v \in X_{0,0}^{T,s}$, then for any $\phi \in H^s(R)$ and $f \in L^1([-T,T]; H^s(R))$, the IVP (1.11) has a unique solution $u \in X_{l,r}^{T,s}$ and

$$\|u\|_{X_{l,r}^{T,s}} \le \beta\left(\|v\|_{X_{0,0}^{T,s}}\right)\left(\|\phi\|_s + \int_{-T}^T \|f(.,t)\|_s dt\right). \tag{1.12}$$

The estimate (1.12) will be key in obtaining differentiability of the map K. The proof of the well-posedness of the IVP (1.1) in the space $H^s(R)$ with $s > 3/4$ is presented in §3. We show that the map K is infinitely many times Frechet differentiable from $H^s(R)$ to $X_{l,r}^{T,s}$ in §4. The analyticity of the map K from $H^s(R)$ to $X_{l,r}^{T,s}$ in the case in which $a(u)$ is a polynomial is established in §5.

Notation. The norm in $L^2(R)$ will be denoted by $\|.\|$ and the norm in $H^s(R)$ will be denoted by $\|.\|_s$. The notation $\|.\|_\infty$ is used to denote the norm in $L^\infty(R)$.

$D^s = (-\partial_x^2)^{s/2}$ and $J^s = (1 - \partial_x^2)^{s/2}$ denote the Riesz and the Bessell potential of order s, respectively.

$[A, B] = AB - BA$, where A, B are operators. Thus $[J^s; f]g = J^s(fg) - fJ^s g$ where f is regarded as a multiplication operator.

$H^\infty(R) := \cap_{s>0} H^s(R)$.

For $1 \le p, q \le \infty$, and $f : R \times [-T, T] \to R$,

$$\|f\|_{L_T^q L_x^p} = \left(\int_{-T}^{T} \left(\int_{-\infty}^{\infty} |f(x,t)|^p dx \right)^{\frac{q}{p}} dt \right)^{\frac{1}{q}}$$

and

$$\|f\|_{L_x^p L_T^q} = \left(\int_{-\infty}^{\infty} \left(\int_{-T}^{T} |f(x,t)|^q dt \right)^{\frac{p}{q}} dx \right)^{\frac{1}{p}}.$$

2. Linear estimates. Let $\{W(t)\}_{-\infty}^{+\infty}$ be the unitary group generated by the linear third order operator $Af = -f'''$ in the space $L^2(R)$. Then the solution of the IVP associated with

$$\begin{cases} \partial_t v + \partial_x^3 v = 0 & \text{for } x,\ t \in R, \\ v(x,0) = v_0(x) \end{cases} \tag{2.1}$$

is given by

$$v(t) = W(t)v_0,$$

and the solution of the inhomogeneous equation

$$\begin{cases} \partial_t v + \partial_x^3 v = f(x,t), & x,\ t \in R, \\ v(x,0) = 0 \end{cases} \tag{2.2}$$

is represented by

$$v(t) = \int_0^t W(t-\tau) f(.,\tau) d\tau.$$

LEMMA 2.1. *For any $s \ge 0$,*

$$\left(\sup_x \int_{-\infty}^{\infty} |D^s \partial_x W(t) v_0|^2 dt \right)^{1/2} \le c \|v_0\|_s \tag{2.3}$$

and

$$\left(\int_{-\infty}^{\infty} \|D^{s+1/4}W(t)v_0\|_{\infty}^4 dt\right)^{1/4} \le c\|v_0\|_s. \tag{2.4}$$

In addition, if $s > 3/4$, *then*

$$\left(\int_{-\infty}^{+\infty} \sup_{[-T,T]} |J^l W(t)v_0|^2(x)dx\right)^{1/2} \le c(1+T)^{\rho}\|v_0\|_s, \tag{2.5}$$

where $l \in [0, s-3/4)$ *and* ρ *is a fixed constant larger than* $3/4$.

Proof. See Kenig, Ponce, and Vega [17, Lem. 2.1, Thm. 2.4, and Cor. 2.9]. □

LEMMA 2.2. *For any* $s \ge 0$ *and* $T > 0$,

$$\|W(t)v_0\|_s = \|v_0\|_s \tag{2.6}$$

and

$$\sup_{[-T,T]} \left\| \int_0^t W(t-\tau)f(.,\tau)d\tau \right\|_s \le \int_{-T}^{T} \|f(.,\tau)\|_s d\tau. \tag{2.7}$$

Proof. (2.6) and (2.7) follow easily from Kato ([13, Lem. 3.1]). □

LEMMA 2.3. *For any* $s \ge 0$ *and* $T > 0$,

$$\left\| D_x^s \partial_x \int_0^t W(t-\tau)f(.,\tau)d\tau \right\|_{L_x^{\infty}L_T^2} \le c\|f\|_{L^1([-T,T];H^s(R))} \tag{2.8}$$

and

$$\left\| D^{s+\frac{1}{4}} \int_0^t W(t-\tau)f(.,\tau)d\tau \right\|_{L_x^{\infty}L_T^4} \le c\int_{-T}^{T} \|f(.,\tau)\|_s d\tau. \tag{2.9}$$

If $s > 3/4$, *then*

$$\left\| J^l \int_0^t W(t-\tau)f(.,\tau)d\tau \right\|_{L_x^2 L_T^{\infty}} \le c(1+T)^{\rho}\int_{-T}^{T} \|f(.,\tau)\|_s d\tau, \tag{2.10}$$

where $l \in [0, s-3/4)$ *and* ρ *is a fixed constant larger than* $3/4$.

Proof. The proof follows from Lemma 2.1 by using Minkowski's integral inequality (cf. [25]). □

LEMMA 2.4. *Let* $s > 1/2$ *and* $T > 0$ *be given. Then there is a constant* $c > 0$ *such that*

$$\int_{-T}^{T} \|u\partial_x v\|_s dt \le cT^{1/2}(1+T)^{\rho}\|u\|_{X_{0,0}^{T,s}}\|v\|_{X_{0,0}^{T,s}} \tag{2.11}$$

and

$$\int_{-T}^{T} \|\partial_x(uv)\|_s dt \le cT^{1/2}(1+T)^{\rho}\|u\|_{X_{0,0}^{T,s}}\|v\|_{X_{0,0}^{T,s}} \tag{2.12}$$

for any $u,\ v \in X_{0,0}^{T,s}$.

Proof. For the proof see Kenig, Ponce, and Vega [18, the claim (4.10)] or Lemma 2.4 in [25]. □

LEMMA 2.5. *Let $s > 0$ be given. Then there is a constant $c > 0$ such that for any $y_k \in H^s(R)$, $k = 1, 2, \ldots, m$,*

$$\left\| \prod_{k=1}^{m} y_k \right\|_s \leq c^m \left(\sum_{j=1}^{m} \|y_j\|_s \prod_{k=1,\, k\neq j}^{m} \|y_k\|_\infty \right) \tag{2.13}$$

and if $s > 1/2$, then

$$\left\| \prod_{k=1}^{m} y_k \right\|_s \leq c^m \prod_{k=1}^{m} \|y_k\|_s \tag{2.14}$$

for any $m \geq 2$, where the constant c in (2.14) may be different from the c in (2.13).

Proof. (2.13) with $m = 2$ is Lemma X4 of Kato and Ponce [14]. The general case follows easily by induction. The proof is complete. □

LEMMA 2.6. *Let $b \in C^\infty(R; R)$ with $b(0) = 0$. Then*

$$\|b(u)\|_s \leq \tilde{b}(\|u\|_s), \qquad s > 1/2,$$

where $\tilde{b}(.)$ is a monotone increasing function depending only on b.

Proof. See Kato ([13, Lem. A.3]). □

LEMMA 2.7. *Let $s > 1/2$ and $T > 0$ be given and assume $a \in C^\infty(R; R)$ with $a(0) = 0$. Then there is a constant $c > 0$ such that for any $u \in X_{0,0}^{T,s}$ and $y \in X_{0,0}^{T,s}$,*

$$\int_{-T}^{T} \|\partial_x(a(u)y)\|_s dt \leq c\beta\left(\|u\|_{X_{0,0}^{T,s}}\right) \|y\|_{X_{0,0}^{T,s}}, \tag{2.15}$$

where $\beta(.)$ is a continuous monotone increasing function only depending on a.

Proof. First of all,

$$\|\partial_x(a(u)y)\|_s \leq \|a'(u)y\partial_x u\|_s + \|a(u)\partial_x y\|_s$$

and it is easy to see by using Lemmas 2.5 and 2.6 that

$$\begin{aligned} \|a'(u)y\partial_x u\|_s &\leq \|(a'(u) - a'(0))y\partial_x u\|_s + |a'(0)|\|y\partial_x u\|_s \\ &\leq c\{\|a'(u) - a'(0)\|_s + |a'(0)|\}\, \|y\partial_x u\|_s \\ &\leq c\beta_1(\|u\|_s)\|y\partial_x u\|_s, \end{aligned}$$

where $\beta_1(.) : R^+ \to R^+$ is a continuous monotone increasing function only depending on a. Using Lemma 2.4 yields

$$\begin{aligned} \int_{-T}^{T} \|a'(u)y\partial_x u\|_s dt &\leq c \sup_{[-T,T]} \beta_1(\|u\|_s) \int_{-T}^{T} \|y\partial_x u\|_s dt \\ &\leq cT^{1/2}(1+T)^\rho \beta_1\left(\|u\|_{X_{0,0}^{T,s}}\right) \|u\|_{X_{0,0}^{T,s}} \|y\|_{X_{0,0}^{T,s}}. \end{aligned}$$

It follows from Kenig, Ponce, and Vega [17, Lem. 2.10] that

$$\begin{aligned} \|a(u)\partial_x y\|_s &= \|J^s(a(u)\partial_x y)\| \\ &= \|a(u)D^s\partial_x y + a(u)(J^s - D^s)\partial_x y + [J^s; a(u)]\partial_x y\| \\ &\leq \|a(u)D^s\partial_x y\| + \|a(u)\|_\infty \|y\|_s \\ &\qquad + c\{\|\partial_x y\|_\infty \|a(u)\|_s + \|a'(u)\partial_x u\|_\infty \|y\|_s\}. \end{aligned}$$

Note that

$$\int_{-T}^{T} \|a(u)\|_\infty \|y\|_s dt \le \sup_{[-T,T]} \|a(u)\|_s \int_{-T}^{T} \|y\|_s dt$$
$$\le cT^{1/2}(1+T)^\rho \beta_2(\|u\|_{X_{0,0}^{T,s}}) \|y\|_{X_{0,0}^{T,s}},$$

where $\beta_2(.) : R^+ \to R^+$ is a continuous monotone increasing function depending only on a,

$$\int_{-T}^{T} \|a(u)\|_s \|\partial_x y\|_\infty dt \le \sup_{[-T,T]} \|a(u)\|_s \int_{-T}^{T} \|\partial_x y\|_\infty dt$$
$$\le cT^{1/2}(1+T)^\rho \beta_2(\|u\|_{X_{0,0}^{T,s}}) \|y\|_{X_{0,0}^{T,s}},$$

$$\int_{-T}^{T} \|a'(u)\partial_x u\|_\infty \|y\|_s dt \le \sup_{[-T,T]} (\|a'(u)\|_\infty \|y\|_s) \int_{-T}^{T} \|\partial_x u\|_\infty dt$$
$$\le cT^{1/2}(1+T)^\rho \beta_3(\|u\|_{X_{0,0}^{T,s}}) \|u\|_{X_{0,0}^{T,s}} \|y\|_{X_{0,0}^{T,s}},$$

where

$$\beta_3(r) = \sup_{|\lambda| \le r} |a'(\lambda)|,$$

and

$$\int_{-T}^{T} \|a(u) D^s \partial_x y\| dt \le T^{1/2} \int_{-T}^{T} \int_R |a(u) D^s \partial_x u|^2 dx dt$$
$$\le \int_R \int_{-T}^{T} |u|^2 |D^s \partial_x u|^2 dx dt \sup_{[-T,T]} \left\| \frac{a(u)}{u} \right\|_\infty$$
$$\le cT^{1/2}(1+T)^\rho \beta_4(\|u\|_{X_{0,0}^{T,s}}) \|u\|_{X_{0,0}^{T,s}} \|y\|_{X_{0,0}^{T,s}},$$

where

$$\beta_4(r) = \sup_{|\lambda| \le r} \left| \frac{a(\lambda)}{\lambda} \right|.$$

We conclude that

$$\int_{-T}^{T} \|\partial_x (a(u)y)\|_s dt \le cT^{1/2}(1+T)^\rho \beta(\|u\|_{X_{0,0}^{T,s}}) \|y\|_{X_{0,0}^{T,s}}$$

for some constant $c > 0$, where

$$\beta(r) = \max \{ r\beta_1(r),\ \beta_2(r),\ r\beta_3(r),\ r\beta_4(r) \}.$$

The proof is complete. □

From the proof of the above lemma we may draw the following corollary.

COROLLARY 2.1. *Let $s > 1/2$ and $T > 0$ be given and assume $a \in C^\infty(R; R)$ with $a'(0) = 0$. Then for any $u \in X_{0,0}^{T,s}$,*

$$\int_{-T}^{T} \|\partial_x (a(u))\|_s dt \le c\beta \left(\|u\|_{X_{0,0}^{T,s}} \right) T^{1/2}(1+T)^\rho \|u\|_{X_{0,0}^{T,s}}, \tag{2.16}$$

where $\beta(.)$ is a continuous monotone increasing function only depending on a.

LEMMA 2.8. *Let $s > 1/2$, $T > 0$ and $a \in C^\infty(R; R)$ be given. Then there exists a constant $c > 0$ such that for any $u \in X_{0,0}^{T,s}$ and $y_k \in X_{0,0}^{T,s}$ with $k = 1, 2, \ldots, m$ and $m \geq 2$,*

$$\int_{-T}^{T} \|\partial_x(a(u)\prod_{k=1}^{m} y_k)\|_s dt \leq c^m \beta(\|u\|_{X_{0,0}^{T,s}}) \prod_{k=1}^{m} \|y_k\|_{X_{0,0}^{T,s}}, \tag{2.17}$$

where $\beta(.)$ is a continuous monotone increasing function only depending on a.

Proof. By applying (2.14) we have

$$\begin{aligned}
\left\|\partial_x\left(a(u)\prod_{k=1}^{m} y_k\right)\right\|_s &\leq \left\|a'(u)\partial_x u \prod_{k=1}^{m} y_k\right\|_s + \sum_{k=1}^{m}\left\|\prod_{j=1,\, j\neq k}^{m} a(u) y_j \partial_x y_k\right\|_s \\
&\leq c^{m+1}\left(\|a'(u) - a'(0)\|_s + |a'(0)|\right)\|y_1\partial_x u\|_s \prod_{k=2}^{m}\|y_k\|_s \\
&\quad + c^{m+1}\left(\|a(u) - a(0)\|_s + |a(0)|\right) \\
&\quad * \left(\sum_{k=2}^{m}\|y_1\partial_x y_k\|_s \prod_{j=2,\, j\neq k}^{m}\|y_j\|_s + \|y_2\partial_x y_1\|_s \prod_{j=3}^{m}\|y_j\|_s\right).
\end{aligned}$$

Then, using Lemmas 2.6 and 2.4 leads to

$$\begin{aligned}
&\int_{-T}^{T}\left\|\partial_x\left(a(u)\prod_{k=1}^{m} y_k\right)\right\|_s dt \leq c_1^{m+1}\sup_{[-T,T]}\left(\|a'(u) - a'(0)\|_s + |a'(0)|\right) \\
&\quad * \prod_{k=2}^{m}\sup_{[-T,T]}\|y_k\|_s \int_{-T}^{T}\|y_1\partial_x u\|_s dt + c_1^{m+1}\sup_{[-T,T]}\left(\|a(u) - a(0)\|_s + |a(0)|\right) \\
&\quad * \left(\sum_{k=2}^{m}\sup_{[-T,T]}\prod_{j=2,\, j\neq k}^{m}\|y_j\|_s \int_{-T}^{T}\|y_1\partial_x y_k\|_s dt + \prod_{j=3}^{m}\sup_{[-T,T]}\|y_j\|_s \sup_{[-T,T]}\|y_2\partial_x y_1\|_s\right) \\
&\leq c_1^{m+1}\beta_1(\|u\|_{X_{0,0}^{T,s}})T^{1/2}(1+T)^\rho\|u\|_{X_{0,0}^{T,s}}\prod_{k=1}^{m}\|y_k\|_{X_{0,0}^{T,s}} \\
&\qquad + mc_1^{m+1}\beta_1(\|u\|_{X_{0,0}^{T,s}})T^{1/2}(1+T)^\rho\prod_{k=1}^{m}\|y_k\|_{X_{0,0}^{T,s}} \\
&\leq c^m\beta(\|u\|_{X_{0,0}^{T,s}})\prod_{k=1}^{m}\|y_k\|_{X_{0,0}^{T,s}}
\end{aligned}$$

for some $c > 0$. The proof is complete. □

Now we turn to consider the linear problem

$$\begin{cases} \partial_t u + \partial_x(a(v)u) + \partial_x^3 u = f(x,t), & x,\, t \in R, \\ u(x,0) = \phi(x), \end{cases} \tag{2.18}$$

where $a(.) \in C^\infty(R; R)$ with $a(0) = 0$.

THEOREM 2.1. *Let $s > 3/4$, $(l,r) \in [0, s-3/4] \times [0, s-3/4)$, $T > 0$ and $v \in X_{0,0}^{T,s}$ be given. Then for any $f \in L^1([-T,T]; H^s(R))$ and $\phi \in H^s(R)$, there exists a unique solution $u \in X_{l,r}^{T,s}$ to (2.18) such that*

$$\|u\|_{X_{l,r}^{T,s}} \le \beta(\|v\|_{X_{0,0}^{T,s}}) \left(\|\phi\|_s + \int_{-T}^{T} \|f(.,t)\|_s dt \right), \tag{2.19}$$

where β is a continuous monotone increasing function only depending on a.

Proof. We use the same contraction principle argument that Kenig, Ponce, and Vega used in [18].

For any given $\phi \in H^s(R)$ and $f \in L^1([-T,T]; H^s(R))$, the inhomogeneous problem

$$\begin{cases} \partial_t u + \partial_x^3 u = f - \partial_x(a(v)w), \\ u(x,0) = \phi(x) \end{cases} \tag{2.20}$$

defines a map

$$\Phi : w \to \text{ the solution } u$$

for any

$$w \in S_b^T = \left\{ w \in X_{0,0}^{T,s} | \quad \Lambda_{0,0}^s(T; w) \le b \right\},$$

where $b > 0$ is to be determined.

Rewrite (2.20) in its integral equation form as follows:

$$u(t) = W(t)\phi + \int_0^t W(t-\tau) \left(f - \partial_x(a(v)w) \right)(.,\tau) d\tau. \tag{2.21}$$

Applying (2.3)–(2.10) and (2.15) to (2.21) leads to

$$\begin{aligned} \Lambda_{l,r}^s(t; u) &\le c \left(\|\phi\|_s + \int_{-t}^{t} \|f(.,\tau)\|_s d\tau \right) + c \int_{-t}^{t} \|\partial_x(a(v)w)\|_s d\tau \\ &\le c \left(\|\phi\|_s + \int_{-T}^{T} \|f(.,\tau)\|_s d\tau \right) + ct^{1/2}(1+t)^{\rho} \beta(\|v\|_{X_{0,0}^{T,s}}) \Lambda_{0,0}^s(t; w) \end{aligned} \tag{2.22}$$

for $(l,r) \in [0, s-3/4] \times [0, s-3/4)$. In particular,

$$\Lambda_{0,0}^s(t; u) \le c \left(\|\phi\|_s + \int_{-T}^{T} \|f\|_s d\tau \right) + c\beta(\|v\|_{X_{0,0}^{T,s}}) t^{1/2}(1+t)^{\rho} \Lambda_{0,0}^s(t; w).$$

Choosing

$$b = 2 \left(\|\phi\|_s + \int_{-T}^{T} \|f\|_s d\tau \right) \tag{2.23}$$

and $0 < T^* < T$ such that

$$cT^*(1+T^*)^{\rho} \beta(\|v\|_{X_{0,0}^{T,s}}) = 1/2, \tag{2.24}$$

we obtain

$$\Lambda^s_{0,0}(T^*;u) \le b. \tag{2.25}$$

Thus, Φ is a map from $S_b^{T^*}$ to $S_b^{T^*}$ with b given by (2.23). The contraction property of the map Φ follows similarly. As a result, there exists a unique $u \in S_b^{T^*}$ such that

$$u(t) = W(t)\phi - \int_0^t W(t-\tau)\left(\partial_x(a(v)u)\right)(\tau)d\tau + \int_0^t W(t-\tau)f(.,\tau)d\tau \tag{2.26}$$

for $-T^* < t < T^*$ and

$$\Lambda^s_{l,r}(T^*;u) \le c\left(\|\phi\|_s + \int_{-T}^{T}\|f\|_s d\tau\right). \tag{2.27}$$

Note that T^* determined by (2.24) only depends on $\beta(\|v\|_{X^{T,s}_{0,0}})$, not on f and ϕ particularly. Thus, a standard argument shows that T^* can be extended to $T^* = T$ and (2.27) holds for $T^* = T$ with another c depending only on $\|v\|_{X^{T,s}_{0,0}}$ and a. The proof is complete. □

Remark 2.1. Theorem 2.1 is still true if $a(v)$ in (2.18) is replaced by

$$\int_0^1 a(\lambda v_1 + (1-\lambda)v_2)d\lambda$$

with $v_1,\ v_2 \in X^{T,s}_{0,0}$.

3. Well-posedness. From now on we assume that $a(u)$ in (1.1) satisfies $a(0) = a'(0) = 0$.

THEOREM 3.1. *Let $s > 3/4$ and $(l,r) \in [0, s-3/4] \times [0, s-3/4)$ be given. Then the following hold:*

(i) *For any $\phi \in H^s(R)$, there exists a $T > 0$ depending only on $\|\phi\|_s$ and a unique solution $u \in C([-T,T]; H^s(R))$ to the* IVP (1.1) *satisfying*

$$\left(\int_{-T}^{T}\|\partial_x u(.,t)\|^4_\infty dt\right)^{1/4} < \infty.$$

Moreover,

$$\|u\|_{X^{T,s}_{l,r}} \le \eta(\|\phi\|_s),$$

where $\eta(.)$ is a continuous monotone increasing function with $\eta(0) = 0$.

(ii) *For any $T' < T$, there exists a neighborhood U of ϕ in $H^s(R)$ such that the map*

$$K : \phi \to u(.,t)$$

from U to $X^{T',s}_{l,r}$ is Lipschitz continuous.

THEOREM 3.2. *Let $s \ge 1$ be given. Then Theorem* 3.1 *is true with T arbitrarily large provided that*

$$\|\phi\|_1 \le \gamma_a,$$

where γ_a is the ceiling of $a = a(u)$ defined by Kato [13, *Thm* 4.4].

Remark 3.1. $\gamma_a = \infty$ if $a'(u) < 0$ or $a'(u) \leq |u|^p$ with $p \leq 3$ as $|u| \to \infty$ (see Kato [13]).

Remark 3.2. Theorem 3.2 follows from Theorem 3.1 and the global a priori estimates for solutions of the IVP (1.1) of Kato [13] by a standard argument.

We start to prove Theorem 3.1 by establishing the following a priori estimate for solutions of (1.1).

PROPOSITION 3.1. *Let $s > 3/4$, $T > 0$, and $(l, r) \in [0, s - 3/4] \times [0, s - 3/4)$ be given. If $u \in X_{0,0}^{T,s}$ is a solution of* (1.1), *then there exists a $T_0 > 0$ depending only on $\|\phi\|_s$ such that*

$$\Lambda_{l,r}^s(T_0; u) < c\|\phi\|_s, \tag{3.1}$$

where $c > 0$ is a constant independent of u.

Proof. Consider the integral equation form of (1.1),

$$u(t) = W(t)\phi - \int_0^t W(t-\tau)\partial_x(a(u))(\tau)d\tau. \tag{3.2}$$

Applying (2.3)–(2.10) to (3.2) yields

$$\Lambda_{l,r}^s(t; u) \leq c\|\phi\|_s + c\int_{-t}^t \|\partial_x(a(u))(\tau)\|_s d\tau \tag{3.3}$$

for any $t \leq T$. In particular, using (2.16) we have

$$\begin{aligned} \Lambda_{0,0}^s(t; u) &\leq c\|\phi\|_s + c\int_{-t}^t \|\partial_x(a(u))(\tau)\|_s d\tau \\ &\leq c\|\phi\|_s + c\beta(\Lambda_{0,0}^s(t; u))t^{1/2}(1+t)^\rho \Lambda_{0,0}^s(t; u). \end{aligned} \tag{3.4}$$

Since $\beta(\Lambda_{0,0}^s(t; u))$ is a continuous increasing function of t, there exists a $t = T_0$ such that

$$c\beta(\Lambda_{0,0}^s(T_0; u))T_0^{1/2}(1+T_0)^\rho = 1/2. \tag{3.5}$$

Then it is from (3.4) that

$$\Lambda_{0,0}^s(T_0; u) \leq 2c\|\phi\|_s. \tag{3.6}$$

Thus the following inequality must hold:

$$c\beta(2c\|\phi\|_s)T_0^{1/2}(1+T_0)^\rho \geq 1/2.$$

It implies that $T_0 > M_1 > 0$ for $M_1 > 0$ depending only on $\|\phi\|_s$. The estimate (3.1) then follows from (3.3), (2.16), and (3.6). The proof is complete. □

PROPOSITION 3.2. *Let $s > 3/4$ and $(l, r) \in [0, s - 3/4] \times [0, s - 3/4)$ be given. For any $\phi \in H^s(R)$, suppose $\phi_\epsilon \in H^\infty$, $\epsilon \in (0, 1)$, with*

$$\lim_{\epsilon \to 0} \phi_\epsilon = \phi_0 \qquad \text{in } H^s(R).$$

Then there exists $T > 0$ such that for any $\epsilon \in (0, 1)$, (1.1) *has a unique solution $u_\epsilon \in C([-T, T]; H^\infty(R))$ with $u_\epsilon(x, 0) = \phi_\epsilon(x)$ satisfying*

$$\Lambda_{l,r}^s(T; u_\epsilon) < c\|\phi_\epsilon\|_s \tag{3.7}$$

for $0 < \epsilon < 1$, where $c > 0$ is independent of ϵ.

Proof. According to Kato [13], $u_\epsilon \in C([-T_\epsilon, T_\epsilon]; H^\infty(R))$, where T_ϵ only depends on $\|\phi_\epsilon\|_s$. Since $\|\phi_\epsilon\|_s$ is uniformly bounded for $\epsilon \in (0,1)$, we may assume that

$$T_\epsilon > T_1 \qquad \text{for any } \epsilon \in (0,1)$$

for some $T_1 > 0$, and therefore

$$u_\epsilon \in X_{0.0}^{T_1,s}.$$

It follows from Proposition 3.1 that

$$\Lambda_{l,r}^s(T; u_\epsilon) \le c\|\phi_\epsilon\|_s$$

for some $T > 0$ and $c > 0$ independent of ϵ. The proof is complete. □

Proof of Theorem 3.1. For $\phi \in H^s(R)$, choose $\phi_\epsilon \in H^\infty(R)$ such that

$$\lim_{\epsilon \to 0} \phi_\epsilon = \phi \qquad \text{in } H^s(R).$$

By Proposition 3.2, there exists a $T > 0$ and $c > 0$ such that for any $\epsilon \in (0,1)$, the IVP

$$\begin{cases} \partial_t u_\epsilon + \partial_x(a(u_\epsilon)) + \partial_x^3 u_\epsilon = 0, \\ \\ u_\epsilon(x,0) = \phi_\epsilon(x) \end{cases}$$

has a unique solution $u_\epsilon \in C([-T,T]; H^\infty(R))$ satisfying

$$\Lambda_{l,r}^s(T; u_\epsilon) < c\|\phi_\epsilon\|_s.$$

We show that u_ϵ is a Cauchy sequence in $X_{l,r}^{T,s}$. Then its limit u as $\epsilon \to 0$ is the desired solution of the IVP (3.1) corresponding to the initial value ϕ.

Let $\epsilon' < \epsilon$ and

$$w = u_\epsilon - u_{\epsilon'}.$$

Then w solves

$$\begin{cases} \partial_t w + \partial_x(A(u_\epsilon, u_{\epsilon'})w) + \partial_x^3 w = 0, \\ \\ w(x,0) = \phi_\epsilon - \phi_{\epsilon'}, \end{cases}$$

where

$$A(u,v) = \int_0^1 a'\left(\lambda u - (1-\lambda)(v)\right) d\lambda.$$

According to Theorem 2.1 and Remark 2.1,

$$\Lambda_{l,r}^s(T; w) \le c^* \|\phi_\epsilon - \phi_{\epsilon'}\|_s$$

for some $c^* > 0$ independent of ϵ, and therefore u_ϵ is a Cauchy sequence in $X_{l,r}^{T,s}$.

Finally, if $T' < T$, there exists a neighborhood U of ϕ in $H^s(R)$ such that the IVP (1.1) defines a map: $\phi \to u(.,t)$ from U to $X_{l,r}^{T',s}$. For any $\psi_1,\ \psi_2 \in U$, let u and v be the solutions of the IVP (1.1) with $u(x,0) = \psi_1(x)$ and $v(x,0) = \psi_2(x)$, respectively. Then, similarly, we have

$$\Lambda_{l,r}^s(T'; u - v) \le c\|\phi - \psi\|_s,$$

where c depends only on $\|\psi_j\|_s$, $j = 1,2$. Therefore the map $\phi \to u$ is Lipschitz continuous. The proof is complete. □

4. Differentiability. Let $s > 3/4$ and $(l, r) \in [0, s-3/4] \times [0, s-3/4)$ be given. According to Theorem 3.1, for any $\phi \in H^s(R)$, there exist a $T > 0$ and a neighborhood U of ϕ in $H^s(R)$ such that (1.1) defines a nonlinear map K from U to $X_{l,r}^{T,s}$,

$$u := K(\psi)$$

for any $\psi \in U$, where u is the solution of (1.1).

Suppose that the map K is n times Frechet differentiable; then its nth order derivative $K^{(n)}(\psi)$ at $\psi \in U$ is a symmetric n-linear map from $H^s(R)$ to $X_{l,r}^{T,s}$ and for any $h_1, \dots, h_n \in H^s(R)$,

$$K^{(n)}(\psi)[h_1, \dots, h_n] = \left\{ \frac{\partial^n}{\partial\xi_1 \dots \partial\xi_n} K\left(\psi + \sum_{k=1}^{n} \xi_k h_k\right) \right\}_{0,\dots,0}.$$

As for the homogeneous polynomial function $K^{(n)}(\psi)[h^n]$ of degree n induced by $K^{(n)}(\psi)$, it is given by

$$K^{(n)}(\psi)[h^n] = \left\{ \frac{d^n}{d\xi^n} K(\psi + \xi h) \right\}_{\xi=0}$$

for any $h \in H^s(R)$.

Let

$$w_{[1,\dots,n]}^{(n)} = K^{(n)}(\psi)[h_1, \dots, h_n] \quad \text{and} \quad y_n = K^{(n)}(\psi)[h^n].$$

Then direct computation shows that $w_{[1,\dots,n]}^{(n)}$ solves

$$\text{(4.1)} \qquad \begin{cases} \partial_t w_{[1]}^{(1)} + \partial_x(a'(u) w_{[1]}^{(1)}) + \partial_x^3 w_{[1]}^{(1)} = 0, \\ w_{[1]}^{(1)}(x, 0) = h_1 \end{cases}$$

for $n = 1$ and

$$\text{(4.2)} \qquad \begin{cases} \partial_t w_{[1,\dots,n]}^{(n)} + \partial_x(a'(u) w_{[1,\dots,n]}^{(n)}) + \partial_x^3 w_{[1,\dots,n]}^{(n)} = -\partial_x(H_n), \\ w_{[1,\dots,n]}^{(n)}(x, 0) = 0 \end{cases}$$

for $n \geq 2$ with $u = K(\psi)$ and

$$\text{(4.3)} \qquad H_n = \sum_{j=2}^{n} \frac{a^{(j)}(u)}{j!} \sum_{k_1+\dots+k_j=n} \sum_{\mathcal{U}} w_{[i_1^1,\dots,i_{k_1}^1]}^{(k_1)} w_{[i_1^2,\dots,i_{k_2}^2]}^{(k_2)} \cdots w_{[i_1^j,\dots,i_{k_j}^j]}^{(k_j)},$$

where $\sum_{\mathcal{U}}$ is the summation over all $(i_1^1, \dots, i_{k_1}^1, \dots i_1^j, \dots, i_{k_j}^j)$ satisfying

$$1 \leq i_1^m < i_2^m < \cdots < i_{k_m}^m \leq n$$

for $m = 1, 2, \dots, j$ and

$$\bigcup_{m=1}^{j} \bigcup_{l=1}^{k_m} \{i_l^m\} = \{1, 2, \dots, n\}.$$

One can easily see that $w^{(1)}_{[k]}$, k=2,...,n, solves (4.1) with h_1 replaced by h_k. Similarly, $w^{(2)}_{[i_1,i_2]}$, $1 \le i_1,\ i_2 \le n$, solves (4.1)–(4.2) with h_1 and h_2 replaced by h_{i_1} and h_{i_2} respectively.

As for y_n, it solves

$$\begin{cases} \partial_t y_1 + \partial_x(a'(u)y_1) + \partial_x^3 y_1 = 0, \\ y_1(x,0) = h \end{cases} \tag{4.4}$$

for $n = 1$ and

$$\begin{cases} \partial_t y_n + \partial_x(a'(u)y_n) + \partial_x^3 y_n = -\partial_x(M_n), \\ y_n(x,0) = 0 \end{cases} \tag{4.5}$$

for $n \ge 2$, where

$$M_n = \sum_{j=2}^{n} \frac{a^{(j)}(u)}{j!} \sum_{k_1+\cdots+k_j=n} \frac{n!}{k_1!\dots k_j!} y_{k_1}\cdots y_{k_j}.$$

On the other hand, according to Theorem 2.1, for given $u \in X^{T,s}_{0,0}$ with $s > 3/4$, (4.1) defines a linear map $\mathcal{K}^{(1)}(u)$ from $H^s(R)$ to $X^{T,s}_{l,r}$ with $(l,r) \in [0, s-3/4] \times [0, s-3/4)$,

$$\mathcal{K}^{(1)}(u)[h_k] = w^{(1)}_{[k]},$$

where $w^{(1)}_{[k]}$ is the solution of (4.1) with the initial value $h_k \in H^s(R)$. Inductively, (4.1)–(4.2) defines an n-linear map $\mathcal{K}^{(n)}(u)$ from the n-fold space $(H^s(R))^n$ to $X^{T,s}_{l,r}$ for any $n \ge 2$.

PROPOSITION 4.1. *Let $s > 3/4$, $(l,r) \in [0, s-3/4] \times [0, s-3/4)$, $T > 0$ and $u \in X^{T,s}_{0,0}$ be given. Then for any $h_1, \dots, h_n \in H^s(R)$, (4.1)–(4.2) has a unique solution $w^{(n)}_{[1,\dots,n]}$, which defines an n-linear map $\mathcal{K}^{(n)}(u)$ from the n-fold space $(H^s(R))^n$ to $X^{T,s}_{l,r}$. Moreover,*

$$\|\mathcal{K}^{(n)}(u)[h_1,\dots,h_n]\|_{X^{T,s}_{l,r}} \le c(n, \|u\|_{X^{T,s}_{0,0}}) \prod_{k=1}^{n} \|h_k\|_s \tag{4.6}$$

for any $n \ge 1$ and $h_1, \dots, h_n \in H^s(R)$ where $c(n,.)$ is a continuous monotone increasing function from R^+ to R^+ with $c(n,0) = 0$.

Proof. It follows from Theorem 2.1 that system (4.1)–(4.2) defines an n-linear map $\mathcal{K}^{(n)}(u)$. We prove estimate (4.6) by induction.

It is easy to see that (4.6) is true with $n = 1$ by applying estimate (2.19) to equation (4.1). Suppose that (4.6) is true for $1 \le k \le N-1$. Applying Lemma 2.8 and estimate (2.19) to equation (4.2) with $n = N$, we obtain

$$\begin{aligned} \|w^{(N)}_{[1,\dots,N]}\|_{X^{T,s}_{l,r}} &\le \beta(\|u\|_{X^{T,s}_{0,0}}) \int_{-T}^{T} \|\partial_x H_N\|_s dt \\ &\le \beta(\|u\|_{X^{T,s}_{0,0}}) \sum_{j=2}^{N} \frac{c^j}{j!} \beta_j(\|u\|_{X^{T,s}_{0,0}}) \sum_{k_1+\cdots+k_j=N} \prod_{l=1}^{j} \|w^{(k_l)}_{[i^l_1,\dots,i^l_{k_l}]}\|_{X^{T,s}_{0,0}} \end{aligned}$$

$$\leq \beta(\|u\|_{X_{0,0}^{T,s}}) \sum_{j=2}^{N} \frac{c^j}{j!} \beta_j(\|u\|_{X_{0,0}^{T,s}}) \sum_{k_1+\cdots+k_j=N} \prod_{l=1}^{j} c(k_l, \|u\|_{X_{0,0}^{T,s}}) \|h_{[i_1^l,\ldots,i_{k_l}^l]}\|_s$$

$$:= c(N, \|u\|_{X_{0,0}^{T,s}}) \prod_{k=1}^{N} \|h_k\|_s.$$

The proof is complete. □

COROLLARY 4.1. *Let $s > 3/4$, $(l, r) \in [0, s - 3/4] \times [0, s - 3/4)$, $T > 0$, and $u \in X_{0,0}^{T,s}$ be given. Then (4.4)–(4.5) defines a homogeneous polynomial $\mathcal{K}^{(n)}(u)[h^n]$ of degree n from $H^s(R)$ to $X_{l,r}^{T,s}$ and*

$$\|\mathcal{K}^{(n)}(u)[h^n]\|_{X_{l,r}^{T,s}} \leq c(n, \|u\|_{X_{0,0}^{T,s}}) \|h\|_s^n \tag{4.7}$$

for any $h \in H^s(R)$.

Now we may formally define the nth Taylor polynomial $P_n(h)$ of the map K at $\phi \in H^s(R)$ as

$$P_n(h) = K(\phi) + \sum_{k=1}^{n} \frac{1}{k!} \mathcal{K}^{(k)}(u)[h^k]$$

$$:= u + \sum_{k=1}^{n} \frac{1}{k!} y_k,$$

where $u = K(\phi)$.

PROPOSITION 4.2. *Let z_n denote the nth Taylor remainder of K at $\phi \in H^s(R)$:*

$$z_n = K(\phi + h) - P_n(h).$$

Then z_n solves

$$\begin{cases} \partial_t z_0 + \partial_x(F_1(u, v) z_0) + \partial_x^3 z_0 = 0, \\ z_0(x, 0) = h(x) \end{cases} \tag{4.8}$$

for $n = 0$ and

$$\begin{cases} \partial_t z_n + \partial_x(F_1(u, v) z_n) + \partial_x^3 z_n = -\partial_x(D_n), \\ z_n(x, 0) = 0 \end{cases} \tag{4.9}$$

for $n \geq 1$, where

$$u = K(\phi), \qquad v = K(\phi + h),$$

and

$$D_n = \sum_{m=2}^{n+1} F_m(u, v) \sum_{k=0}^{n+1-m} z_k \sum_{k_1+\cdots+k_{m-1}=n-k} q_{k_1} \cdots q_{k_{m-1}} \tag{4.10}$$

with

$$F_m(u, v) = \int_0^1 \lambda_1^{m-1} \ldots \int_0^1 \lambda_{m-1} \int_0^1 a^{(m)} \left(\prod_{j=1}^{m} \lambda_j v + \left(1 - \prod_{j=1}^{m} \lambda_j \right) u \right) d\lambda_m \ldots d\lambda_1$$

and

$$q_m = \frac{y_m}{m!}$$

for $m = 1, \ldots, n+1$.

Proof. Direct computation shows easily that (4.8) and (4.9) with $n = 1$ are true. Assume that (4.9) is true for $n \leq N$. For $n = N+1$, by definition,

$$z_{N+1} = z_N - \frac{1}{(N+1)!} y_{N+1} = z_N - q_{N+1},$$

where q_{N+1} solves

$$\begin{cases} \partial_t q_{N+1} + \partial_x (a'(u) q_{N+1}) + \partial_x^3 q_{N+1} = -\partial_x (E_{N+1}), \\ q_{N+1} = 0 \end{cases} \tag{4.11}$$

with

$$E_{N+1} = \sum_{m=2}^{N+1} \frac{a^{(m)}(u)}{m!} \sum_{k_1 + \cdots + k_m = N+1} q_{k_1} \cdots q_{k_m}.$$

Hence

$$\partial_t z_{N+1} + \partial_x^3 z_{N+1} = -\partial_x \left(F_1(u,v) z_N - a^{(1)}(u) q_{N+1} \right) - \partial_x (G_{N+1}),$$

where

$$\begin{aligned} G_{N+1} = & \sum_{m=2}^{N+1} F_M(u,v) \sum_{k=0}^{N+1-m} z_k \sum_{k_1 + \cdots + k_{M_1} = N-k} q_{k_1} \cdots q_{k_{m-1}} \\ & - \sum_{m=1}^{N+1} \frac{a^{(m)}(u)}{m!} \sum_{k_1 + \cdots + k_m = N+1} q_{k_1} \cdots q_{k_m}. \end{aligned}$$

We can readily verify that

$$\begin{aligned} & F_1(u,v) z_N - a'(u) q_{N+1} - G_{N+1} \\ & \quad = F_1(u,v) z_{N+1} + F_2(u,v) z_0 q_{N+1} \\ & \qquad + \sum_{m=2}^{N+1} F_m(u,v) \sum_{k=1}^{N+2-m} z_k \sum_{k_1 + \cdots + k_{m-1} = N+1-k} q_{k_1} \cdots q_{k_{m-1}} \\ & \qquad + \sum_{m=3}^{N+2} F_m(u,v) z_0 \sum_{k_1 + \cdots + k_{m-1} = N+1} q_{k_1} \cdots q_{k_{m-1}} \\ & \quad = F_1(u,v) z_{N+1} + \sum_{m=2}^{N+2} F_m(u,v) \sum_{k=0}^{N+2-m} z_k \sum_{k_1 + \cdots + k_{m-1} = N+1-k} q_{k_1} \cdots q_{k_{m-1}}. \end{aligned}$$

We conclude that

$$\begin{cases} \partial_t z_{N+1} + \partial_x (F_1(u,v) z_{N+1}) + \partial_x^3 z_{N+1} = -\partial_x (D_{N+1}), \\ z_{N+1}(x,0) = 0. \end{cases}$$

The proof is completed by induction. □

THEOREM 4.1. *Let $s > 3/4$ and $(l, r) \in [0, s-3/4] \times [0, s-3/4)$ be given. Then, for any $\phi^* \in H^s(R)$, there exist a $T > 0$ and a neighborhood U of ϕ^* in $H^s(R)$ such that the nonlinear map K defined by the* IVP (1.1) *is infinitely many times Frechet differentiable in U from $H^s(R)$ to $X_{l,r}^{T,s}$. Its nth derivative $K^{(n)}$ at $\psi \in U$ is given by*

$$K^{(n)}(\psi)[h_1, \dots, h_n] = \mathcal{K}^{(n)}(u)[h_1, \dots, h_n]$$

for any $h_1, \dots, h_n \in H^s(R)$, where $\mathcal{K}^{(n)}(u)$ is defined by system (4.1)–(4.2) *with $u = K(\psi)$.*

Proof. We only need to prove that for any $\psi \in U$,

$$K(\phi + h) = \sum_{k=0}^{n} \frac{1}{k!} \mathcal{K}^{(k)}(\phi)[h^k] + o(\|h\|_s^n)$$

as $h \to 0$ in $H^s(R)$ uniformly for $\|\phi - \psi\|_s \le \|h\|_s$ by the converse Taylor theorem (see [7]).

Let

$$v = K(\phi + h), \qquad u = K(\phi), \qquad y_{(k)} = \mathcal{K}^{(k)}(u)[h^k]$$

for $1 \le k \le n$ and

$$z_0 = v - u, \qquad z_1 = z_0 - y_0, \qquad z_n = z_{n-1} - \frac{1}{n!} y_n.$$

Choose $\delta_1 > 0$ such that

$$S_{\delta_1}(\psi) = \{\phi \in H^s(R),\ \|\phi - \psi\|_s \le \delta_1\} \subset U.$$

By Corollary 4.1, for any $\phi \in S_{\delta_1}(\psi)$,

$$\|q_k\|_{X_{l,r}^{T,s}} \le c(k, \|u\|_{X_{0,0}^{T,s}}) \|h\|_s^k, \qquad k = 1, 2, \dots, n,$$

where $c(k, \|u\|_{X_{0,0}^{T,s}})$ is uniformly bounded on $S_{\delta_1}(\psi)$.

It suffices to prove that

$$\|z_n\|_{X_{l,r}^{T,s}} \le \gamma_n \|h\|_s^{n+1} \tag{4.12}$$

for $n \ge 0$, where γ_n is uniformly bounded for $\phi \in S_{\delta_1}(\psi)$.

We prove estimate (4.12) by induction. First it is easy to check that (4.12) is true for $n = 0$ by applying Theorem 2.1 to equation (4.8). Suppose (4.12) is true for $n \le N$; then applying Theorem 2.1 and Lemma 2.8 to (4.9) with $n = N + 1$, we have that

$$\begin{aligned}
&\|z_{N+1}\|_{X_{l,r}^{T,s}} \\
&\quad \le c \sum_{m=2}^{N+2} \sum_{k=0}^{N+2-m} \sum_{k_1 + \cdots + k_{m-1} = N+1-k} \int_{-T}^{T} \|\partial_x \left(F_m(u, v) z_k q_{k_1} \dots q_{k_{m-1}}\right)\|_s dt \\
&\quad \le c \sum_{m=2}^{N+2} \sum_{k=0}^{N+2-m} \sum_{k_1 + \cdots + k_{m-1} = N+1-k} c^m \beta_m \Lambda_{0,0}^s(T; z_k) \prod_{j=1}^{m-1} \Lambda_{0,0}^s(T; q_{k_j}) \\
&\quad \le c \sum_{m=2}^{N+1} \sum_{k=0}^{N+2-m} \sum_{k_1 + \cdots + k_{m-1} = N+1-k} c^m \gamma_m \beta_m \prod_{j=1}^{m-1} c(k_j, \|u\|_{X_{0,0}^{T,s}}) \|h\|_s^{N+2} \\
&\quad := \gamma_{N+1} \|h\|_s^{N+2}
\end{aligned}$$

The proof is complete. □

COROLLARY 4.2 (Taylor's Formula). *For any $\phi \in U$ and $h \in H^s(R)$, if*

$$\phi + \xi h \in U \qquad \text{for any } \xi \in (0,1),$$

then

$$K(\phi+h) = \sum_{j=0}^{n-1} \frac{1}{j!} K^{(j)}(\phi)[h^j] + \int_0^1 \frac{(1-\xi)^{n-1}}{n!} K^{(n)}(\phi+\xi h)[h^n] d\xi$$

for any $n \geq 1$.

Proof. See [7, Thm. 8.14.3]. □

5. Analyticity. To show further that the map K is analytic, i.e., it has Taylor series expansion at any $\phi \in H^s(R)$,

$$K(\phi+h) = \sum_{n=0}^{\infty} \frac{K^{(n)}(\phi)}{n!}[h^n], \tag{5.1}$$

we need more accurate estimates of the nth derivative $y_n = K^{(n)}(\phi)[h^n]$ and the nth Taylor remainder z_n.

PROPOSITION 5.1. *Let $s > 3/4$, $T > 0$, $(l,r) \in [0, s-3/4] \times [0, s-3/4)$ and $u \in X_{0,0}^{T,s}$ be given. If y_n is the solution of (4.4)–(4.5), then there exists a sequence $\alpha(n)$ given by*

$$\alpha(1) = 1 \tag{5.2}$$

and

$$\alpha(n) = \sum_{j=2}^{n} \frac{\beta_j}{j!} \sum_{k_1+\cdots+k_j=n} \alpha(k_1)\alpha(k_2)\ldots\alpha(k_j) \tag{5.3}$$

for $n \geq 2$ such that

$$\|y_n\|_{X_{l,r}^{T,s}} \leq c^n n! \alpha(n) \|h\|_s^n \tag{5.4}$$

for any $n \geq 1$, where $c > 0$ is a constant independent of n and h,

$$\beta_j = \beta_j(\|u\|_{X_{0,0}^{T,s}})$$

is a continuous monotone function depending only on $a^{(j)}$, and $\beta_j \equiv 0$ if $a^{(j)} \equiv 0$.

Proof. Note that $q_n = y_n/n!$ solves

$$\begin{cases} \partial_t q_1 + \partial_x(a'(u)q_1) + \partial_x^3 q_1 = 0, \\ q_1(x,0) = h(x) \end{cases} \tag{5.5}$$

for $n = 1$ and

$$\begin{cases} \partial_t q_n + \partial_x(a'(u)q_n) + \partial_x^3 q_n = -\partial_x \left(\sum_{j=2}^{n} \frac{a^{(j)}(u)}{j!} \sum_{k_1+\cdots+k_j=n} q_{k_1}\cdots q_{k_j} \right), \\ q_n(x,0) = 0 \end{cases} \tag{5.6}$$

for $n \geq 2$.

Applying (2.19) to (5.5) yields

$$\|q_1\|_{X_{l,r}^{T,s}} \leq c\|h\|_s.$$

Assume that

$$\|q_m\|_{X_{l,r}^{T,s}} \leq c^{2m-1}\alpha(m)\|h\|_s^m \qquad \text{for } 1 \leq m \leq N.$$

Then applying (2.19) to (5.7) with $n = N+1$, we obtain

$$\begin{aligned}
\|q_{N+1}\|_{X_{l,r}^{T,s}} &\leq c\sum_{j=2}^{N+1}\frac{1}{j!}\sum_{k_1+\cdots+k_j=N+1}\int_{-T}^{T}\|\partial_x(a^{(j)}(u)q_{k_1}\dots q_{k_j})\|_s dt \\
&\leq c\sum_{j=2}^{N+1}\frac{\beta_j}{j!}\sum_{k_1+\cdots+k_j=N+1}\prod_{l=1}^{j}\|q_{k_l}\|_{X_{0,0}^{T,s}} \\
&\leq c\sum_{j=2}^{N+1}\frac{\beta_j}{j!}\sum_{k_1+\cdots+k_j=N+1}\prod_{l=1}^{j}c^{2k_l-1}\|h\|_s^{k_l}\alpha(k_l) \\
&\leq c^{2(N+1)-1}\alpha(N+1)\|h\|_s^{N+1}.
\end{aligned}$$

Thus we have proved by induction that

$$\|y_n\|_{X_{l,r}^{T,s}} \leq c^{2n-1}n!\alpha(n)\|h\|_s^n,$$

which is (5.4) with a different $c > 0$. The proof is completed by induction. □

Similarly, we have the following proposition.

PROPOSITION 5.2. *Let $s > 3/4$, $T > 0$, $(l,r) \in [0, s-3/4] \times [0, s-3/4)$, and $u, v \in X_{0,0}^{T,s}$ be given. If z_n is the solution of* (4.8)–(4.9) *with y_n being the solution of* (4.4)–(4.5), *then there exists a sequence $\gamma(n)$ given by*

$$\gamma(0) = \eta_1, \tag{5.7}$$

$$\gamma(n) = \eta_1\sum_{m=2}^{n+1}\eta_m\sum_{k=0}^{n+1-m}\gamma(k)\sum_{k_1+\cdots+k_{m-1}=n-k}\prod_{i=1}^{m-1}c^i\alpha(i) \tag{5.8}$$

for $n \geq 1$ with $\alpha(n)$ given by (5.2)–(5.3) *such that*

$$\|z_n\|_{X_{l,r}^{T,s}} \leq \gamma(n)\|h\|_s^{n+1}, \qquad n \geq 0, \tag{5.9}$$

where

$$\eta_j = \eta_j(\|u\|_{X_{0,0}^{T,s}}, \|v\|_{X_{0,0}^{T,s}}),$$

depending only on $a^{(j)}$, is a continuous function of $\|u\|_{X_{0,0}^{T,s}}$ and $\|v\|_{X_{0,0}^{T,s}}$ with $\eta_j \equiv 0$ if $\alpha^{(j)} \equiv 0$, and $c > 0$ is a constant independent of n and h.

Now we turn to consider convergence of the Taylor series (5.1).

PROPOSITION 5.3. *Let $\phi \in U$. If there is a $\delta_1 > 0$ and $c > 0$ such that*

$$S_{\delta_1}(\phi) = \{\psi \in H^s(R) | \quad \|\phi - \psi\|_s < \delta_1\} \subset U$$

and

$$\gamma(n) \leq c^n, \qquad n \geq 0 \tag{5.10}$$

uniformly for $\psi \in S_{\delta_1}(\phi)$ where $\gamma(n)$ is defined by (5.7)–(5.8) *with $u = K(\phi)$ and $v = K(\psi)$, then there exists a $\delta > 0$ such that series* (5.1) *converges in $X_{l,r}^{T,s}$ uniformly for $\|h\|_s \leq \delta$.*

Proof. Consider the nth Taylor remainder

$$z_n = K(\phi + h) - \sum_{j=0}^{n} \frac{1}{j!} K^{(j)}(\phi)[h^j].$$

According to Proposition 5.2 and hypothesis (5.10),

$$\|z_n\|_{X_{l,r}^{T,s}} \leq \gamma(n)\|h\|_s^{n+1} \leq c^n \|h\|_s^{n+1},$$

where $c > 0$ is independent of n and $h \in H^s(R)$ with $\|h\|_s < \delta_1$. Choose $\delta > 0$ such that

$$\delta < \frac{1}{2c};$$

then

$$\|z_n\|_{X_{l,r}^{T,s}} \leq (1/2)^n, \qquad n \geq 1$$

for any $h \in H^s(R)$ with $\|h\|_s \leq \delta$. The proof is complete. □

In the following we show that if $a(u)$ in (1.1) is a polynomial of degree N,

$$a(u) = \sum_{j=2}^{N} b_j u^j, \tag{5.11}$$

with b_j, $j = 1, 2, \ldots, N$ being real constants, then hypothesis (5.10) is satisfied for any $\phi \in U$. Consequently, the map K is an analytic map from $U \subset H^s(R)$ to $X_{l,r}^{T,s}$.

First we prove a technical lemma.

LEMMA 5.1. *Let $N \geq 1$ be a given integer and α_n be a sequence given by*

$$\alpha_N(1) = 1, \tag{5.12}$$

$$\alpha_N(n) = \sum_{j=2}^{n} \frac{b_j}{j!} \sum_{k_1 + \cdots + k_j = n} \alpha_N(k_1) \ldots \alpha_N(k_j) \quad \text{for } 2 \leq n \leq N-1, \tag{5.13}$$

and

$$\alpha_N(n) = \sum_{j=2}^{N} \frac{b_j}{j!} \sum_{k_1 + \cdots + k_j = n} \alpha_N(k_1) \ldots \alpha_N(k_j) \quad \text{for } n \geq N. \tag{5.14}$$

Then there exists a constant $c > 0$ such that

$$\alpha_N(n) \le c^n \tag{5.15}$$

for any $n \ge 1$.

Remark 5.1. In the case $N = 2$ and $b_2 = 1$, the lemma is Proposition 3.4 in [25] where it is shown that

$$\alpha_2(n) = \frac{2^{n-1}(2n-3)!!}{n!} \qquad \text{for any } n \ge 2.$$

Proof of Lemma 5.1. Note that $\alpha_N(n)$, for any $n \ge 1$, is uniquely determined by (5.12)–(5.15) inductively. In particular, one may obtain

$$\alpha_N(1),\ \alpha_N(2), \dots, \alpha_N(N-1)$$

explicitly by computation.

Let

$$\begin{aligned} P_N(x,y) = y &- \sum_{j=2}^{N} \frac{b_j}{j!} \left(\sum_{k=1}^{N-1} \alpha_N(k) x^k + y \right)^j \\ &- \sum_{j=2}^{N} \frac{b_j}{j!} \sum_{k=j}^{N-1} x^k \sum_{k_1+\cdots+k_j=k} \alpha_N(k_1) \dots \alpha_N(k_j). \end{aligned} \tag{5.16}$$

It is easy to check that

$$P_N(0,0) = 0, \qquad \frac{\partial}{\partial y} P_N(0,0) = 1.$$

By the implicit function theorem, the equation

$$P_N(x,y) = 0 \tag{5.17}$$

has a unique solution

$$y = f(x) \qquad \text{with } f(0) = 0$$

defined in a neighborhood of $x = 0$ such that

$$P_N(x, f(x)) = 0 \qquad \text{for any } |x| \le \delta$$

with some $\delta > 0$. Moreover, $f^{(j)}(0) = 0$ for $j = 1, 2, \dots, N-1$.

In fact, the function $y = f(x)$ is a real analytic function in a neighborhood of $x = 0$. In particular, it has a Taylor series expansion at $x = 0$,

$$f(x) = \sum_{j=N}^{\infty} d_j x^j, \tag{5.18}$$

which is uniformly convergent for $|x| < \delta$ with some $\delta > 0$.

To see this, let

$$g = \sum_{k=1}^{N-1} \alpha_N(k) x^k + y$$

and

$$z = \sum_{k=1}^{N-1} \alpha_N(k)x^n + \sum_{j=2}^{N} \frac{b_j}{j!} \sum_{k=j}^{N-1} x^k \sum_{k_1+\cdots+k_j=k} \alpha_N(k_1)\ldots\alpha_N(k_j) := h(x).$$

Then, equation (5.17) may be written as

$$g - \sum_{j=2}^{N} \frac{b_j}{j!} g^j = z,$$

which obviously has an analytic solution $g(z)$ in a neighborhood of $z = 0$ such that $g(0) = 0$. Thus $y = f(x) = g(h(x)) - \sum_{k=1}^{N-1} \alpha_N(k)x^k$ is an analytic function in a neighborhood of $x = 0$, since $h(x)$ is a polynomial function of x.

Now let us define

$$d_j = \alpha_N(j) \qquad \text{for } j = 1, 2, \ldots, N-1.$$

It follows from (5.17) and (5.18) by direct computation that

$$\begin{aligned}\sum_{j=N}^{\infty} d_j x^j &= \sum_{j=2}^{N} \frac{b_j}{j!} \left(\sum_{k=1}^{N-1} \alpha_N(k)x^k + \sum_{k=N}^{\infty} d_k x^k \right)^j \\ &\quad - \sum_{j=2}^{N} \frac{b_j}{j!} \sum_{k=j}^{N-1} x^k \sum_{k_1+\cdots+k_j=k} \alpha_N(k_1)\ldots\alpha_N(k_j) \\ &= \sum_{k=N}^{\infty} \left(\sum_{j=2}^{N} \frac{b_j}{j!} \sum_{k_1+\cdots+k_j=k} d_{k_1}\ldots d_{k_j} \right) x^k\end{aligned}$$

for any $|x| < \delta$. Hence

$$d_j = \alpha_N(j), \qquad j = 1, 2, \ldots, N-1$$

and

$$d_k = \sum_{j=2}^{N} \frac{b_j}{j!} \sum_{k_1+\cdots+k_j=k} d_{k_1}\ldots d_{k_j}$$

for any $k \geq N$. That is, for $j \geq 1$ d_j also satisfies the induction relation (5.12)–(5.15). By uniqueness,

$$\alpha_N(k) = d_k \qquad \text{for any } k \geq 1.$$

Furthermore, note that $\{d_k\}_N^\infty$ are coefficients of Taylor series (5.18). Therefore there exists a $c > 0$ such that

$$\alpha_N(n) = d_n \leq c^n \qquad \text{for all } n \geq 1.$$

The proof is complete. □

THEOREM 5.1. *Let $s > 3/4$ and $(l, r) \in [0, s - 3/4] \times [0, s - 3/4)$ be given and suppose that the function $a(u)$ in (1.1) is a polynomial function of degree N . Then*

for any $\phi \in H^s(R)$, there is a $T > 0$ and a neighborhood U of ϕ in $H^s(R)$ such that the IVP (1.1) *defines an analytic map K from U to $X_{l,r}^{T,s}$, i.e., for any $\psi \in U$, there is a $\delta > 0$ such that the Taylor series*

$$K(\psi + h) = K(\psi) + \sum_{n=1}^{\infty} \frac{1}{n!} K^{(n)}(\psi)[h^n]$$

uniformly converges for $\|h\|_s \leq \delta$ in the space $X_{l,r}^{T,s}$.

Proof. Since U is an open subset in $H^s(R)$, there is a $\delta_1 > 0$ such that if $h \in H^s(R)$ with $\|h\|_s \leq \delta_1$, then

$$\psi + h \in U.$$

Denote

$$v = K(\psi + h), \qquad u = K(\psi)$$

and

$$z_0 = v - u, \qquad z_n = z_{n-1} - \frac{1}{n!} y_n \quad \text{for } n \geq 1.$$

Then, by Propositions 5.1 and 5.2,

$$\|y_n\|_{X_{l,r}^{T,s}} \leq n! c^n \alpha(n) \|h\|_s^n$$

and

$$\|z_n\|_{X_{l,r}^{T,s}} \leq \gamma(n) \|h\|_s^{n+1}.$$

Note that $a^{(j)}(u) \equiv 0$ for $j \geq N + 1$ by the assumption. Then $\alpha(n)$ is given by

$$\alpha(1) = 1,$$

$$\alpha(n) = \sum_{j=2}^{n} \frac{\beta_j}{j!} \sum_{k_1 + \cdots + k_j = n} \alpha(k_1) \ldots \alpha(k_j)$$

for $2 \leq n \leq N - 1$ and

$$\alpha(n) = \sum_{j-2}^{N} \frac{\beta_j}{j!} \sum_{k_1 + \cdots + k_j = n} \alpha(k_1) \ldots \alpha(k_j)$$

for $n \geq N$. As for $\gamma(n)$,

$$\gamma(0) = \eta_1,$$

$$\gamma(n) = \eta_1 \sum_{j=2}^{n} \eta_j \sum_{k=0}^{n+1-j} \gamma(k) \sum_{k_1 + \cdots + k_{j-1} = n-k} \prod_{i=1}^{j-1} c^{k_i} \alpha(k_i)$$

for $1 \le n \le N-1$, and

$$\gamma(n) = \eta_1 \sum_{j=2}^{N} \eta_j \sum_{k=0}^{n+1-j} \gamma(k) \sum_{k_1+\cdots+k_{j-1}=n-k} \prod_{i=1}^{j-1} c^{k_i} \alpha(k_i).$$

According to Proposition 5.3, it suffices to show that there is a $c_* > 0$ such that

$$\gamma(n) \le c_*^n$$

for any $n \ge N$ with c_* independent of n and $\|h\|_s \le \delta_1$. To this end, we first see from Lemma 5.1 that

$$\alpha(n) \le c_1^n, \qquad n \ge N$$

with some $c_1 > 0$ independent of n and h. Thus, for $0 \le j \le N$,

$$\sum_{k_1+\cdots+k_{j-1}=n-k} \prod_{i=1}^{j-1} c^{k_i} \alpha(k_i) \le c_2^{n-k}$$

for some $c_2 > 0$ independent of n and

$$\begin{aligned} \gamma(n) &\le c_3 \sum_{j=2}^{N} \sum_{k=0}^{n+1-j} \gamma(k) \sum_{k_1+\cdots+k_{j-1}=n-k} \prod_{i=1}^{j-1} c^{k_i} \alpha(k_i) \\ &\le \sum_{k=0}^{n-1} c_4^{n-k} \gamma(k) \end{aligned}$$

for some $c_4 > 0$ independent of n. Then we can readily verify by induction that

$$\gamma(n) \le 2^{n-1} c_4^n \le c_*^n$$

for some $c_* > 0$ independent of n and h. The proof is complete. □

COROLLARY 5.1. *Assume that $a(u)$ in (1.1) is a polynomial and $s \ge 1$. Let*

$$Y_a^s := H^s(R) \cap \left\{ \phi \in H^1(R),\ \|\phi\|_1 \le \gamma_a \right\},$$

where γ_a is the ceiling of a (cf. Thm. 3.2). Then for any $T > 0$, the map K defined by the IVP *(1.1) is analytic from Y_a^s to $X_{l,r}^{T,s}$ with any $(l, r) \in [0, s-3/4] \times [0, s-3/4)$.*

REFERENCES

[1] J. L. BONA AND R. SCOTT, *Solutions of the Korteweg–de Vries equation in fractional order Sobolev spaces*, Duke Math. J., 43 (1976), pp. 87–99.

[2] J. L. BONA AND R. SMITH, *The initial value problem for the Korteweg–de Vries equation*, Roy. Soc. London Ser. A, 278 (1978), pp. 555–601.

[3] J. L. BONA AND B.-Y. ZHANG, *The initial value problem for the forced Korteweg–de Vries equation*, IMA preprint 1253, University of Minnesota, 1994.

[4] J. BOURGAIN, *Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, Part* I*: Schrödinger Equations*, Geom. Funct. Anal., 3 (1993), pp. 107–156.

[5] ———, *Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, Part* II*: The* KdV *Equation*, Geom. Funct. Anal., 3 (1993), pp. 209–262.

[6] P. CONSTANTIN AND J. C. SAUT, *Local smoothing properties of dispersive equations*, J. Amer. Math. Soc., 1 (1988), pp. 413–446.
[7] J. DIEUDONNÉ, *Foundation of Modern Analysis*, Academic Press, New York, 1969.
[8] D. HENRY, *Geometric Theory of Semilinear Parabolic Equation*, Lecture Notes in Math. 840, Springer-Verlag, New York, Berlin, Heidelberg, 1981.
[9] W. CRAIG, T. KAPPELER, AND W. A. STRAUSS, *Gain of regularity for equations of* KdV *type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 147–186.
[10] J. GINIBRE AND Y. TSUTSUMI, *Uniqueness for the generalized Korteweg–de Vries equations*, SIAM J. Math. Anal., 20 (1989), pp. 1388–1425.
[11] J. GINIBRE, Y. TSUTSUMI, AND G. VELO, *Existence and uniqueness of solutions for the generalized Korteweg-de Vries equation*, Math Z., 203 (1990), pp. 9–36.
[12] T. KATO, *Quasilinear equations of evolutions, with applications to partial differential equation*, Lecture Notes in Math. 448, Springer-Verlag, New York, Berlin, Heidelberg, 1975, pp. 27–50.
[13] ———, *On the Cauchy problem for the (generalized) Korteweg–de Vries equations*, Adv. Math. Supplementary Studies, Studies Appl. Math., 8 (1983), pp. 93–128.
[14] T. KATO AND G. PONCE, *Commutator estimates and the Euler and Navier–Stokes equations*, Comm. Pure Appl. Math., 41 (1988), pp. 897–907.
[15] C. E. KENIG, G. PONCE, AND L. VEGA, *On the (generalized) Korteweg–de Vries equation*, Duke Math. J., 59 (1989), pp. 585–610.
[16] ———, *Oscillatory integral and regularity of dispersive equations*, Indiana Univ. Math. J., 40 (1991), pp. 37–69.
[17] ———, *Well-posedness of the initial value problem for the Korteweg–de Vries*, J. Amer. Math. Soc., 4 (1991), pp. 323–347.
[18] ———, *Well-posedness and scattering results for the generalized Korteweg–de Vries equation via the contraction principle*, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.
[19] ———, *The Cauchy problem for the Korteweg–de Vries equation in Sobolev space of negative indices*, Duke Math. J., 71 (1993), pp. 1–22.
[20] S. N. KRUZHKOV AND A. V. FAMINSKII, *Generalized solutions of the Cauchy problem for the Korteweg–de Vries equation*, Math. USSR-Sb., 48 (1984), pp. 93–138.
[21] G. PONCE AND L. VEGA, *Nonlinear small data scattering for the generalized Korteweg–de Vries equation*, J. Funct. Anal., 90 (1990), pp. 445–457.
[22] J. C. SAUT, *Sur quelques généralizations de l'équations de Korteweg–de Vries*, J. Math. Pures Appl., 58 (1979), pp. 21–61.
[23] J. C. SAUT AND R. TEMAM, *Remarks on the Korteweg–de Vries equation*, Israel J. Math., 24 (1976), pp. 78–87.
[24] L. VEGA, *The Schrödinger equation: Pointwise convergence to the initial data*, Proc. Amer. Math. Soc., 102 (1988), pp. 874–878.
[25] B.-Y. ZHANG, *Taylor series expansion for solutions of the* KdV *equation with respect to their initial values*, J. Funct. Anal., 129 (1995), pp. 293–324.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1514 1526, November 1995

008

CONVERGENCE OF DOUBLE OBSTACLE PROBLEMS TO THE GENERALIZED GEOMETRIC MOTION OF FRONTS*

RICARDO H. NOCHETTO† AND CLAUDIO VERDI‡

Abstract. The connection between the generalized geometric motion of interfaces, interpreted in the viscosity sense, and a singularly perturbed parabolic problem with double obstacle ± 1 and small parameter ε is examined. This approach retains the local character of the limit problem, because the noncoincidence set, where all the action takes place, is a thin transition layer of thickness $\mathcal{O}(\varepsilon)$ irrespective of the forcing term. Zero-level sets are shown to converge past singularities to the generalized motion by mean curvature with forcing, provided no fattening occurs. If the underlying viscosity solution satisfies a nondegeneracy property, namely, its gradient does not vanish, then our results yield interface error estimates and layer width estimates of order $\mathcal{O}(\varepsilon)$. The proofs are based on constructing viscosity subsolutions and supersolutions to the double obstacle problem in terms of the signed distance function and approximate traveling waves dictated by formal asymptotics.

Key words. reaction-diffusion, double obstacle, generalized curvature driven motion, interface convergence, interface error estimates

AMS subject classifications. 35B25, 35K55, 35K57, 35K85, 49Q05

1. Introduction.

In this paper we investigate the relation between the generalized curvature driven motion of interfaces and a singularly perturbed parabolic double obstacle problem. The zero-level set of u_ε, the solution to the singularly perturbed reaction-diffusion partial differential equation (PDE)

$$\varepsilon \partial_t u_\varepsilon - \varepsilon \Delta u_\varepsilon + \frac{1}{2\varepsilon} \Psi'(u_\varepsilon) = \frac{c_0}{2} g \qquad \text{in } \Omega \times (0,T), \tag{1.1}$$

with quartic-like double equal well potential Ψ and $c_0 := \int_{-1}^{1} \sqrt{\Psi(s)}ds$, is known to converge to an interface $\Sigma(t)$ that moves (formally) in the normal direction with velocity

$$V(\mathbf{x},t) = \kappa(\mathbf{x},t) + g(\mathbf{x},t) \qquad \forall\, \mathbf{x} \in \Sigma(t). \tag{1.2}$$

Hereafter $\kappa(\mathbf{x},t)$ indicates the sum of the principal curvatures of $\Sigma(t)$ at $\mathbf{x}$ and g is a forcing term defined in $\Omega \times (0,T)$. This connection has been rigorously established in [3], [9] in the case of no fattening, that is, provided $\Sigma(t)$ has empty interior. Equation (1.1), in turn arises in the Landau–Ginzburg theory of phase transitions [1]. A typical example of potential is the quartic $\Psi(s) = (s^2-1)^2$, but there is no physical reason why Ψ has to be of that form or even smooth. The key condition on Ψ to achieve the geometric law (1.2) in the limit is that the two wells possess equal depth. Exploiting such a freedom of choice, we consider the double obstacle potential

$$\Psi(s) := \begin{cases} 1-s^2 & \text{if } s \in [-1,1], \\ +\infty & \text{if } s \notin [-1,1]. \end{cases} \tag{1.3}$$

* Received by the editors September 14, 1993; accepted for publication (in revised form) March 31, 1994. This work was partially supported by National Science Foundation grant DMS-9305935, the Ministero Università Ricerca Scientifica Tecnologica, and Consiglio Nazionale delle Ricerche contracts 92.00833.01 and 93.00564.01.

† Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

‡ Dipartimento di Matematica, Università di Milano, 20133 Milano, Italy.

In contrast to the usual reaction-diffusion approach with a quartic-like nonlinearity, the coupling of (1.1) and (1.3) exhibits a narrow transition region of order $\mathcal{O}(\varepsilon)$, outside of which $u_\varepsilon = \pm 1$ irrespective of g. Since the resulting problem has to be solved within such a thin noncoincidence set, where all the action takes place, we realize that this approach retains the geometric (or local) structure of the original problem while being insensitive to singularity formation. This property is essential for numerical purposes, and it bears some intrinsic interest as well. There is, however, no complete theoretical justification of this formulation because the *optimal* interface error estimates of [14]–[16] of order $\mathcal{O}(\varepsilon^2)$ are only valid before the onset of singularities, that is provided the motion is smooth. On the other hand, the convergence results of [3], [9], which are valid past singularities, apply to *smooth* quartic-type potentials but not to the (singular) double obstacle (1.3).

In this paper we demonstrate that the zero-level sets $\Sigma_\varepsilon(t)$ of u_ε converge to $\Sigma(t)$ past singularities provided no fattening occurs. The proof is based, as in [3], [9], on constructing viscosity supersolutions to the reaction-diffusion PDE in terms of the signed distance function d to suitable level sets of the viscosity solution ω to the generalized geometric motion (1.2) [7], [10], [12]. The novelties here are the presence of obstacles, which entail lack of regularity of u_ε even for smooth initial data, and the use of an explicit traveling wave dictated by formal asymptotics. Inspired by [19], the first issue is tackled with a suitable notion of viscosity supersolution and corresponding comparison principle. On the other hand, dealing with an approximate explicit traveling wave with fixed transition layer width $\pi\varepsilon$, thereby independent of g, makes the construction of supersolutions simpler than in [3], [9] and certainly more transparent. It also avoids considering generalized flows (1.2) corresponding to perturbed forcing terms $g \pm \mathcal{O}(1)$ as in [3], and yields a new (local) linear rate of convergence $\mathcal{O}(\varepsilon)$ for interfaces, along with layer width estimates, under a further *nondegeneracy* assumption on ω. In fact, if $|\nabla\omega(\mathbf{x},t)| > 0$ for $\mathbf{x} \in \Sigma(t)$, then

$$\operatorname{dist}\big(\mathbf{x}, \Sigma_\varepsilon(t)\big) \le C|\nabla\omega(\mathbf{x},t)|^{-1}\varepsilon.$$

Since this requirement is always valid before the onset of singularities, our results extend those in [6] past singularities. It is remarkable that the linear rate is preserved between singularities, which applies to a number of geometric flows [2], [18].

The local character of the double obstacle formulation, together with its convergence properties even beyond singularities, leads to a robust but effective numerical tool: the *dynamic mesh algorithm* of [13]. This is a finite element solver that solely triangulates the transition layer and then updates it, after having solved the discrete problem, to advance the algorithm in time. Such a simple but crucial idea results in savings of computing time and memory allocation along with enhanced singularity resolution via a space-time dependent relaxation parameter $\varepsilon(\mathbf{x},t)$. This claim has been confirmed both theoretically and numerically [13]–[17] and is further supported here with a rigorous convergence result and error estimates past singularities.

This paper is organized as follows. In §2 we recall key properties of d and introduce the notion of viscosity supersolutions to the double obstacle problem. In §3 we show how to obtain an explicit representation of an approximate traveling wave via formal asymptotics. With the candidate for supersolution at hand, we perform a formal calculation in §4 that presumes regularity of d and confirms the desired properties of our supersolution. This is useful for understanding the main idea behind the quite technical rigorous discussion of §5. A (viscosity) comparison lemma, adequate for obstacle problems, is fully discussed in §6. Both the explicit form of supersolutions

and the comparison principle are key ingredients in proving the convergence of $\Sigma_\varepsilon(t)$ to $\Sigma(t)$ together with interface and layer width estimates; this is carried out in §7.

2. Viscosity solutions. Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain and set $Q = \Omega \times (0,T)$ for $T > 0$. Let the forcing term g satisfy

$$g \in W^{2,1}_\infty(Q). \tag{2.1}$$

Remark 2.1. A weaker condition, meaningful whenever $g = g_\varepsilon$ depends on ε, is written as

$$\varepsilon \|\partial_t g_\varepsilon - \Delta g_\varepsilon\|_{L^\infty(Q)} + \|\nabla g_\varepsilon\|_{L^\infty(Q)} \le C. \tag{2.2}$$

This could happen if g_ε is the solution of another PDE, as in the phase field model for solidification [5], or just a regularized version of a rougher g.

Let $\Sigma_0 \subset \Omega$ be a closed oriented manifold of codimension 1 and class C^1, and let d_0 denote the signed distance function to Σ_0 that is positive outside Σ_0. Let ω denote the (continuous) viscosity solution of the nonlinear degenerate parabolic PDE

$$\partial_t \omega - \Big(\delta_{ij} - \tfrac{\partial_{x_i}\omega\, \partial_{x_j}\omega}{|\nabla\omega|^2}\Big)\partial_{x_i x_j}\omega - g|\nabla\omega| = 0 \qquad \text{in } \mathbb{R}^n \times (0,T), \tag{2.3}$$

satisfying $\omega(\cdot,0) = d_0(\cdot)$ [7], [8], [10], [12]. Such an expression says that level sets of ω evolve formally in their normal direction with velocity $V = \partial_t\omega/|\nabla\omega| = \kappa + g$, as stated in (1.2). Let $\Sigma(t)$ indicate the zero-level set of ω. Since $\Sigma(t)$ is independent of the special form of $\omega(\cdot,0)$, provided $\{\omega(\cdot,0) < 0\} = \text{Interior}\,(\Sigma_0)$, $\Sigma(t)$ is called a generalized geometric evolution for (1.2) [7], [10], [12]. Let $I(t)$ be the inside and $O(t)$ be the outside of $\Sigma(t)$ defined in terms of ω by

$$I(t) := \{\mathbf{x} \in \Omega : \omega(\mathbf{x},t) < 0\}, \qquad O(t) := \{\mathbf{x} \in \Omega : \omega(\mathbf{x},t) > 0\}.$$

Assume $\Sigma(t)$ remains within Ω for all $t \in [0,T]$, whence $\Omega = \Sigma(t) \cup I(t) \cup O(t)$. This condition is guaranteed for $g = 0$ whenever Ω contains the convex hull of Σ_0 [10]. Let $d(\cdot,t)$ denote the signed distance function to the front $\Sigma(t)$ that is positive in $O(t)$; thus $d(\cdot,0) = d_0(\cdot)$. Such a function d satisfies the following property in the viscosity sense [3, p. 446]:

$$\partial_t d - \Delta d - g\big(\mathbf{x} - d(\mathbf{x},t)\nabla d(\mathbf{x},t), t\big) \ge 0 \qquad \text{in } \{d > 0\}, \tag{2.4}$$

at least provided $\Sigma(t)$ has empty interior (no fattening). When $\Sigma(t)$ has a nonempty interior, $d(\mathbf{x},t)$ must be replaced by the distance between $\mathbf{x}$ and $\Omega^-(t) := \{\omega(\cdot,t) < 0\}$ for all $\mathbf{x} \in \Omega\backslash\overline{\Omega^-(t)}$. Equation (2.4) is also valid, but with ≤ 0, for $-\text{dist}\big(\mathbf{x}, \Omega^+(t)\big)$ for all $\mathbf{x} \in \Omega\backslash\overline{\Omega^+(t)}$. In what follows we will use $d(\mathbf{x},t)$ instead of $\text{dist}\big(\mathbf{x}, \Omega^-(t)\big)$ or $-\text{dist}\big(\mathbf{x}, \Omega^+(t)\big)$ for notational simplicity. Inequality (2.4) implies that whenever $\varphi \in C^\infty(Q)$ *is such that* $d - \varphi$ *has a minimum at* $(\mathbf{x}_0,t_0) \in Q$, *where* $d(\mathbf{x}_0,t_0) = \varphi(\mathbf{x}_0,t_0) > 0$, *then*

$$\partial_t\varphi(\mathbf{x}_0,t_0) - \Delta\varphi(\mathbf{x}_0,t_0) - g\big(\mathbf{x}_0 - \varphi(\mathbf{x}_0,t_0)\nabla\varphi(\mathbf{x}_0,t_0), t_0\big) \ge 0. \tag{2.5}$$

In addition, the following regularity properties of d hold [9, p. 1101]:

$$\begin{aligned} &d \text{ is lower semicontinuous in } \{d > 0\}, \text{ upper semicontinuous in } \{d < 0\},\\ &\text{continuous from below in time, and Lipschitz continuous in space.} \end{aligned} \tag{2.6}$$

The variational approximation to $\Sigma(t)$ via a singularly perturbed double obstacle problem is written as follows. We consider the (singular) potential Ψ defined in (1.3)

and let $c_0 = \int_{-1}^{1} \sqrt{\Psi(s)}ds = \frac{\pi}{2}$. The subdifferential Ψ' of Ψ is the maximal graph

$$\tfrac{1}{2}\Psi'(s) := \operatorname{sign}^{-1}(s) - s = \begin{cases} (-\infty, 1] & \text{if } s = -1, \\ -s & \text{if } s \in (-1,1), \\ [-1, +\infty) & \text{if } s = 1. \end{cases}$$

Let u_ε be the solution to the parabolic double obstacle PDE

$$\varepsilon\partial_t u_\varepsilon - \varepsilon\Delta u_\varepsilon - \tfrac{1}{\varepsilon}u_\varepsilon + \operatorname{sign}^{-1}(u_\varepsilon) \ni \tfrac{c_0}{2}g \qquad \text{in } Q, \tag{2.7}$$

subject to the initial and boundary conditions $u_\varepsilon(\mathbf{x}, 0) = \operatorname{sign}\big(d_0(\mathbf{x})\big)$ for $\mathbf{x} \in \Omega$ and $u_\varepsilon = 1$ on $\partial\Omega \times (0,T)$, respectively. Although this problem has both a variational [11] and viscosity interpretation [8], [19], the latter turns out to be more convenient in this setting.

Thus we conclude this section with a definition of viscosity solutions for obstacle problems [8], [19]. We say that a function u_ε^+ *is a viscosity supersolution to the double obstacle problem* (2.7) *if and only if* $u_\varepsilon^+ \geq -1$, *and if* $u_\varepsilon^+ - \varphi$ *attains a minimum at* $(\mathbf{x}_0, t_0) \in Q$ *for* $\varphi \in C^\infty(Q)$ *and* $u_\varepsilon^+(\mathbf{x}_0, t_0) = \varphi(\mathbf{x}_0, t_0) < 1$, *then*

$$\mathcal{J}\varphi := \varepsilon\partial_t\varphi - \varepsilon\Delta\varphi - \tfrac{1}{\varepsilon}\varphi - \tfrac{c_0}{2}g \geq 0 \qquad \text{at } (\mathbf{x}_0, t_0). \tag{2.8}$$

Similarly, we can define a viscosity subsolution. A function u_ε is called a *viscosity solution* of (2.7) if it is both a supersolution and a subsolution.

3. Traveling waves and asymptotics. We study the traveling wave q_α, solution of the following boundary value problem: *given* α, $|\alpha| \ll 1$, *we seek* $v_\alpha, x_\alpha \in \mathbb{R}$ *and* $q_\alpha \in C^{1,1}(\mathbb{R})$ *such that* $|q_\alpha(x)| = 1$ *for all* $x \in \mathbb{R}\backslash(-x_\alpha, x_\alpha)$, $q_\alpha'(x) > 0$ *for all* $x \in (-x_\alpha, x_\alpha)$, *and*

$$q_\alpha''(x) + q_\alpha(x) - v_\alpha q_\alpha'(x) = -\alpha \qquad \text{in } (-x_\alpha, x_\alpha);$$

v_α is the velocity of the traveling wave. it is easy to see that there exists only one v_α,

$$v_\alpha = \tfrac{1}{x_\alpha}\log\big(\tfrac{1+\alpha}{1-\alpha}\big) = \tfrac{2}{c_0}\alpha + \mathcal{O}(\alpha^3),$$

such that $x_\alpha = \pi/2\sqrt{1 - v_\alpha^2/4} > \frac{\pi}{2}$ and the corresponding explicit expression of q_α is

$$q_\alpha(x) = \exp\big(v_\alpha\tfrac{x+x_\alpha}{2}\big)(1-\alpha)\Big(\sin\big(\sqrt{1-v_\alpha^2/4}\;x\big) + \tfrac{v_\alpha}{2\sqrt{1-v_\alpha^2/4}}\cos\big(\sqrt{1-v_\alpha^2/4}\;x\big)\Big) - \alpha.$$

Note that $q_\alpha(\mp x_\alpha) = \mp 1$, $q_\alpha'(\mp x_\alpha) = 0$, $q_\alpha''(\mp x_\alpha^{\pm}) = \pm 1 - \alpha$.

In what follows we will choose $\alpha(\mathbf{x}, t) := \varepsilon\frac{c_0}{2}g(\mathbf{x}, t)$ and denote the corresponding traveling waves by $q_\varepsilon(x; \mathbf{x}, t)$. We stress that $v_{\alpha(\mathbf{x},t)}$ and $x_{\alpha(\mathbf{x},t)}$ depend implicitly on $(\mathbf{x}, t)$ via $g(\mathbf{x}, t)$. An explicit and simpler representation of q_ε can be obtained as follows by means of asymptotics [14], [15], [17].

The unique absolute minimizer γ of the functional $\mathcal{F}(\varphi) := \int_{\mathbb{R}} \big(|\varphi'(x)|^2 + \Psi(\varphi(x))\big)\,dx$, such that $\gamma(0) = 0$, $|\gamma(x)| \leq 1$ in $\mathbb{R}$, and $\lim_{x\to\pm\infty}\gamma(x) = \pm 1$, is given by

$$\gamma(x) := \begin{cases} -1 & \text{if } x < -\frac{\pi}{2}, \\ \sin x & \text{if } x \in [-\frac{\pi}{2}, \frac{\pi}{2}], \\ +1 & \text{if } x > \frac{\pi}{2}. \end{cases}$$

Hence $\gamma \in C^{1,1}(\mathbb{R})$ and solves the (elliptic) double obstacle problem $\gamma''(x) - \frac{1}{2}\Psi'\big(\gamma(x)\big) \ni 0$ in $\mathbb{R}$. In particular,

$$\gamma''(x) + \gamma(x) = 0 \qquad \forall\, x \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}). \tag{3.1}$$

In addition, let us define the function $\eta \in C^{1,1}(\mathbb{R})$,

$$\eta(x) := \begin{cases} \frac{1}{2}\big(x\gamma(x) - c_0 + \gamma'(x)\big) & \text{if } |x| \le \frac{\pi}{2}, \\ 0 & \text{if } |x| > \frac{\pi}{2}, \end{cases}$$

which satisfies $|\eta(x)|, |\eta'(x)| < \gamma'(x)$ in $(-\frac{\pi}{2}, \frac{\pi}{2})$ and solves the problem

$$\eta''(x) + \eta(x) = \gamma'(x) - \tfrac{c_0}{2} \qquad \forall\, x \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}). \tag{3.2}$$

Asymptotics suggest the following approximation Γ of the traveling wave q_ε [14], [15], [17]:

$$\Gamma(x; \mathbf{x}, t) := \gamma(x) + \varepsilon g(\mathbf{x}, t)\eta(x) \qquad \forall\, x \in \mathbb{R},\ (\mathbf{x}, t) \in Q; \tag{3.3}$$

Γ is written without subscript ε for notational simplicity. Note that $\Gamma(\cdot; \mathbf{x}, t) \in C^{1,1}(\mathbb{R})$ is strictly increasing in $(-\frac{\pi}{2}, \frac{\pi}{2})$, because $\Gamma'(x) > \frac{1}{2}\gamma'(x)$ for small ε, and satisfies

$$\Gamma''(x) + \Gamma(x) - \varepsilon g\big(\Gamma'(x) - \tfrac{c_0}{2}\big) = -\varepsilon^2 g^2 \eta'(x) = \gamma'(x)\mathcal{O}(\varepsilon^2) \qquad \text{in } (-\tfrac{\pi}{2}, \tfrac{\pi}{2}), \tag{3.4}$$

as a consequence of (3.1) and (3.2); hereafter we use the notation $\Gamma^{(k)} = \partial_x^k \Gamma$. In view of (2.2), the following (formal) property of Γ holds in Q:

$$\varepsilon(\partial_t \Gamma - \Delta_{\mathbf{x}} \Gamma) - 2\nabla_{\mathbf{x}} \Gamma' \cdot \nabla_{\mathbf{x}} d = \varepsilon^2 \eta(\partial_t g - \Delta_{\mathbf{x}} g) - 2\varepsilon \eta' \nabla_{\mathbf{x}} g \cdot \nabla_{\mathbf{x}} d = \gamma' \mathcal{O}(\varepsilon). \tag{3.5}$$

This illustrates the advantage of having the explicit expression (3.3) and corresponding uniform transition region $(-\frac{\pi}{2}, \frac{\pi}{2})$ independent of $(\mathbf{x}, t) \in Q$. Compare with [3].

4. Supersolution: Formal discussion. We intend to use (2.4) to motivate the construction of supersolutions u_ε^+ to the double obstacle problem (2.7). Set

$$G := \|\nabla g\|_{L^\infty(Q)} + 1 \qquad \text{and} \qquad \sigma(t) := \Lambda e^{2G(T-t)},$$

where $\Lambda > 0$ is to be selected. With Γ as in (3.3), our candidate for supersolution is written as

$$u_\varepsilon^+(\mathbf{x}, t) := \Gamma\big(y(\mathbf{x}, t); \mathbf{x}, t\big) = \gamma\big(y(\mathbf{x}, t)\big) + \varepsilon g(\mathbf{x}, t)\eta\big(y(\mathbf{x}, t)\big) \qquad \forall\, (\mathbf{x}, t) \in Q, \tag{4.1}$$

where y denotes the stretched variable, with time-dependent shift $\frac{\pi}{2} + \sigma(t)$,

$$y(\mathbf{x}, t) := \tfrac{d(\mathbf{x}, t)}{\varepsilon} - \tfrac{\pi}{2} - \sigma(t). \tag{4.2}$$

LEMMA 4.1. *For $\Lambda > 0$ sufficiently large, u_ε^+ is a (formal) supersolution to* (2.7).

Proof. Since no confusion is possible, we set $u := u_\varepsilon^+$. We have to show that $\mathcal{J}u = \varepsilon \partial_t u - \varepsilon \Delta u - \frac{1}{\varepsilon} u - (c_0/2)g \ge 0$. Since

$$\varepsilon \partial_t u = \Gamma' \partial_t d - \varepsilon \sigma'(t)\Gamma' + \varepsilon^2 \eta \partial_t g,$$
$$\varepsilon \Delta u = \tfrac{1}{\varepsilon}\Gamma'' + \Gamma' \Delta d + 2\varepsilon \eta' \nabla g \cdot \nabla d + \varepsilon^2 \eta \Delta g,$$

we have

$$\mathcal{J}u = -\tfrac{1}{\varepsilon}\big(\Gamma'' + \Gamma - \varepsilon g(\Gamma' - \tfrac{c_0}{2})\big) + \Gamma'\big(\partial_t d - \Delta d - g\big) - \varepsilon \sigma'(t)\Gamma' - 2\varepsilon \eta' \nabla g \cdot \nabla d + \varepsilon^2 \eta(\partial_t g - \Delta g).$$

We only have to examine the case $-1 \le u(\mathbf{x}, t) < 1$, that is, $y < \frac{\pi}{2}$. If $y < -\frac{\pi}{2}$, then $\Gamma = -1$ and $\eta = \eta' = \Gamma' = \Gamma'' = 0$, whence $\mathcal{J}u = \frac{1}{\varepsilon} - (c_0/2)g > 0$ for sufficiently small ε. If $-\frac{\pi}{2} < y < \frac{\pi}{2}$ instead, then $0 < d(\mathbf{x}, t) < \varepsilon(\sigma(t) + \pi)$ and

$$\big|g(\mathbf{x}, t) - g\big(\mathbf{x} - d(\mathbf{x}, t)\nabla d(\mathbf{x}, t), t\big)\big| \le \|\nabla g\|_{L^\infty(Q)} d(\mathbf{x}, t) \le \varepsilon G\big(\sigma(t) + \pi\big). \tag{4.3}$$

In view of (2.4), we infer that

$$\Gamma'(\partial_t d - \Delta d - g) \geq -\varepsilon G\Gamma'\big(\sigma(t)+\pi\big).$$

Because of the definition of $\sigma(t)$, (3.4), (3.5) and property $\Gamma' > \frac{1}{2}\gamma'$, we thus have

$$\begin{aligned} \mathcal{J}u &\geq \varepsilon\Gamma'\big(-G\sigma(t) - \sigma'(t) + \mathcal{O}(1)\big) \\ &= \varepsilon\Gamma'\big(G\sigma(t) + \mathcal{O}(1)\big) \geq \varepsilon\Gamma'\big(G\Lambda + \mathcal{O}(1)\big) \geq 0 \end{aligned} \tag{4.4}$$

for a suitable choice of $\Lambda > 0$. Since $\Gamma(\cdot;\mathbf{x},t) \in C^{1,1}(\mathbb{R})$ by construction, we conclude that $\Gamma''\big(y(\mathbf{x},t);\mathbf{x},t\big)$ is a bounded function and, thus, a distribution without mass concentrated on $\big\{y(\mathbf{x},t) = -\frac{\pi}{2}\big\}$. The proof is thus complete. []

Remark 4.1. We see that the term $G\sigma(t)$ that controls $\mathcal{O}(1)$ in (4.4) is the effect of shift σ. A similar effect is obtained in [3] upon considering the distance to a perturbed motion $V = \kappa + g + \mathcal{O}(1)$ instead of (1.2), but with a time-independent shift. This is only feasible in proving convergence without error estimates; see Remark 7.4 below.

Remark 4.2. It is apparent from (4.3) that the exponential form of σ cannot be avoided unless g is independent of $\mathbf{x}$. In such a case, we could consider the linear shift $\sigma(t) := \Lambda(T-t)$ with $\Lambda > 0$ chosen sufficiently large for u_ε^+ to be a formal supersolution again. When $g = 0$, the shift becomes constant, namely, $\sigma(t) = 0$. The above argument in Lemma 4.1 still applies and is much simpler than those in [3], [9].

Remark 4.3. A subsolution can be constructed along the same lines, namely,

$$u_\varepsilon^- := \Gamma\big(\tfrac{d(\mathbf{x},t)}{\varepsilon} + \tfrac{\pi}{2} + \sigma(t);\mathbf{x},t\big) \qquad \forall\ (\mathbf{x},t) \in Q. \tag{4.5}$$

5. Supersolution: Rigorous discussion. We want to prove that $u := u_\varepsilon^+$ defined in (4.1) is a viscosity supersolution of (2.7). Since $u \geq -1$, we only have to demonstrate (2.8). That is, suppose $\varphi \in C^\infty(Q)$ is such that $u - \varphi$ attains a minimum at $(\mathbf{x}_0,t_0) \in Q$ and $u(\mathbf{x}_0,t_0) = \varphi(\mathbf{x}_0,t_0) < 1$. Hence, in view of (4.1) and (4.2),

$$y_0 := y(\mathbf{x}_0,t_0) < \tfrac{\pi}{2}, \qquad \text{that is,} \qquad d(\mathbf{x}_0,t_0) < \varepsilon\big(\sigma(t_0)+\pi\big).$$

We have to show that

$$\mathcal{J}\varphi = \varepsilon\partial_t\varphi - \varepsilon\Delta\varphi - \tfrac{1}{\varepsilon}\varphi - \tfrac{c_0}{2}g \geq 0 \qquad \text{at } (\mathbf{x}_0,t_0), \tag{5.1}$$

which, in turn, we split into several steps.

1. First of all we point out that we can always assume that the minimum of $u-\varphi$ at $(\mathbf{x}_0,t_0)$ is strict, that is,

$$U(\mathbf{x},t) := (u-\varphi)(\mathbf{x},t) > 0 \qquad \text{for } (\mathbf{x},t) \neq (\mathbf{x}_0,t_0). \tag{5.2}$$

In fact, for $0 < \alpha \ll 1$, we have

$$\varphi_\alpha(\mathbf{x},t) := \varphi(\mathbf{x},t) - \alpha\big(|\mathbf{x}-\mathbf{x}_0|^2 + |t-t_0|^2\big) < \varphi(\mathbf{x},t) \qquad \text{for } (\mathbf{x},t) \neq (\mathbf{x}_0,t_0);$$

thus $u - \varphi_\alpha > U \geq 0$. If we obtain (5.1) for φ_α, then

$$\varepsilon\partial_t\varphi - \varepsilon\Delta\varphi - \tfrac{1}{\varepsilon}\varphi - \tfrac{c_0}{2}g \geq 2n\alpha \qquad \text{at } (\mathbf{x}_0,t_0).$$

Now taking $\alpha \downarrow 0$ leads to (5.1) for φ.

2. Let $y_0 < -\frac{\pi}{2}$. By virtue of (4.2) and the last two properties in (2.6), there exists $\alpha > 0$ such that, for all $|\mathbf{x}-\mathbf{x}_0| \leq \alpha$ and $0 \leq t_0 - t \leq \alpha$,

$$y(\mathbf{x},t) < -\tfrac{\pi}{2}, \qquad \text{whence} \qquad -1 \leq u(\mathbf{x},t) = \Gamma\big(y(\mathbf{x},t);\mathbf{x},t\big) \leq -1.$$

The facts that u is constant for all those $(\mathbf{x},t)$ and U attains a minimum at $(\mathbf{x}_0,t_0)$ imply

$$-\Delta\varphi \geq 0, \qquad \partial_t\varphi \geq 0 \qquad \text{at } (\mathbf{x}_0,t_0).$$

Therefore, $\varphi(\mathbf{x}_0,t_0) = -1$ yields

$$\varepsilon\partial_t\varphi - \varepsilon\Delta\varphi - \tfrac{1}{\varepsilon}\varphi - \tfrac{c_0}{2}g \geq \tfrac{1}{\varepsilon} - \tfrac{c_0}{2}g \geq 0 \qquad \text{at } (\mathbf{x}_0,t_0)$$

for ε sufficiently small, and (5.1) follows.

3. Let $y_0 \geq -\frac{\pi}{2}$, that is, $d(\mathbf{x}_0,t_0) \geq \varepsilon\sigma(t_0) > 0$. We first point out that, because of the lower semicontinuity of d (see (2.6)), for $\beta > 0$ sufficiently small, there is a neighborhood of $(\mathbf{x}_0,t_0)$ where $y(\mathbf{x},t) > -\frac{\pi}{2} - \beta$ and $d(\mathbf{x},t) > \varepsilon\big(\sigma(t) - \beta\big) > 0$; in addition, $y_0 < \frac{\pi}{2} - \beta$. We would like to use the fact that d satisfies (2.4). To do so, we must be able to construct a smooth function $\delta(\mathbf{x},t)$ close to $d(\mathbf{x},t)$ via inversion of Γ. Since Γ is not strictly increasing, we define δ as follows. Let γ_α and η_α be regularizations of γ and η by convolution with a smooth kernel ζ_α, whose support is contained in $(-\alpha,\alpha)$, where $0 < \alpha \leq \beta$ is sufficiently small; thus $|\eta_\alpha|, |\eta_\alpha'| \leq \gamma_\alpha'$ in $\mathbb{R}$. Similarly, $g_\alpha \in C^\infty(Q)$ is a smooth approximation of g verifying $\|g_\alpha - g\|_{L^\infty(Q)} \leq C\frac{\alpha}{\varepsilon}$, which is consistent with (2.2). Let

$$\Gamma_\alpha(x;\mathbf{x},t) := \gamma_\alpha(x) + \alpha x + \varepsilon g_\alpha(\mathbf{x},t)\eta_\alpha(x) \tag{5.3}$$

and observe that

$$\Gamma_\alpha'(x;\mathbf{x},t) \geq \tfrac{1}{2}\gamma_\alpha'(x) + \alpha \geq \alpha > 0 \qquad \text{for all } x \in \mathbb{R},\ (\mathbf{x},t) \in Q, \tag{5.4}$$

$$\Gamma_\alpha(x;\mathbf{x},t) \to_{\alpha\downarrow 0} \Gamma(x;\mathbf{x},t) \text{ uniformly for } x \text{ in compact sets of } \mathbb{R},\ (\mathbf{x},t) \in Q. \tag{5.5}$$

Consider the function

$$U_\alpha(\mathbf{x},t) := \Gamma_\alpha\big(y(\mathbf{x},t);\mathbf{x},t\big) - \varphi(\mathbf{x},t).$$

Again using the fact that d is lower semicontinuous in conjunction with (5.2), (5.4), and (5.5), we infer the existence of a point $(\mathbf{x}_\alpha,t_\alpha)$, where U_α attains a local minimum, say ν_α, such that $(\mathbf{x}_\alpha,t_\alpha) \to_{\alpha\downarrow 0} (\mathbf{x}_0,t_0)$ and

$$-\tfrac{\pi}{2} - \beta < y_\alpha := y(\mathbf{x}_\alpha,t_\alpha) < \tfrac{\pi}{2} - \beta. \tag{5.6}$$

The rightmost inequality in (5.6) is a consequence of $y_\alpha \to_{\alpha\downarrow 0} y_0 < \frac{\pi}{2} - \beta$. If this were not true, then there would exist $0 < \xi \leq \beta$ and a subsequence, still labeled y_α, satisfying $y_\alpha \geq y_0 + \xi$ and, in view of the minimality of $(\mathbf{x}_\alpha,t_\alpha)$ and (5.4),

$$\Gamma_\alpha(y_0;\mathbf{x}_0,t_0) - \varphi(\mathbf{x}_0,t_0) = U_\alpha(\mathbf{x}_0,t_0) \geq U_\alpha(\mathbf{x}_\alpha,t_\alpha) \geq \Gamma_\alpha(y_0+\xi;\mathbf{x}_\alpha,t_\alpha) - \varphi(\mathbf{x}_\alpha,t_\alpha).$$

Upon taking $\alpha \downarrow 0$ and using (5.5), $(\mathbf{x}_\alpha,t_\alpha) \to_{\alpha\downarrow 0} (\mathbf{x}_0,t_0)$, $-\frac{\pi}{2} < y_0 + \xi < \frac{\pi}{2}$, and the fact that Γ is strictly increasing in $(-\frac{\pi}{2},\frac{\pi}{2})$, we derive the contradiction

$$\Gamma(y_0;\mathbf{x}_0,t_0) \geq \Gamma(y_0+\xi;\mathbf{x}_0,t_0) > \Gamma(y_0;\mathbf{x}_0,t_0).$$

Since $\Gamma_\alpha(\cdot;\mathbf{x},t)$ is strictly increasing, the following relation uniquely defines $\delta \in C^\infty(Q)$:

$$\varphi(\mathbf{x},t) = \Gamma_\alpha\big(z(\mathbf{x},t);\mathbf{x},t\big) - \nu_\alpha, \qquad \text{where} \qquad z(\mathbf{x},t) = \tfrac{\delta(\mathbf{x},t)}{\varepsilon} - \tfrac{\pi}{2} - \sigma(t).$$

Moreover,

$$\Gamma_\alpha\big(y(\mathbf{x},t);\mathbf{x},t\big) - \varphi(\mathbf{x},t) = U_\alpha(\mathbf{x},t) \geq \nu_\alpha = \Gamma_\alpha\big(z(\mathbf{x},t);\mathbf{x},t\big) - \varphi(\mathbf{x},t),$$

and equality holds at $(\mathbf{x}_\alpha, t_\alpha)$. We thus deduce that $(\mathbf{x}_\alpha, t_\alpha)$ is also a minimum for $d-\delta$ and $0 < d(\mathbf{x}_\alpha, t_\alpha) = \delta(\mathbf{x}_\alpha, t_\alpha)$. Consequently, (2.5) leads to

$$\partial_t\delta(\mathbf{x}_\alpha, t_\alpha) - \Delta\delta(\mathbf{x}_\alpha, t_\alpha) - g\big(\mathbf{x}_\alpha - \delta(\mathbf{x}_\alpha, t_\alpha)\nabla\delta(\mathbf{x}_\alpha, t_\alpha), t_\alpha\big) \geq 0. \tag{5.7}$$

Since $\varphi, \delta \in C^\infty$, the following pointwise calculation makes sense:

$$\begin{aligned}\partial_t\varphi &= \Gamma_\alpha'\big(\tfrac{1}{\varepsilon}\partial_t\delta - \sigma'(t)\big) + \partial_t\Gamma_\alpha,\\ \Delta\varphi &= \tfrac{1}{\varepsilon}\Gamma_\alpha'\Delta\delta + \tfrac{1}{\varepsilon^2}\Gamma_\alpha''|\nabla\delta|^2 + \tfrac{2}{\varepsilon}\nabla_{\mathbf{x}}\Gamma_\alpha'\cdot\nabla\delta + \Delta_{\mathbf{x}}\Gamma_\alpha.\end{aligned}$$

Property $\sigma'(t) = -2G\sigma(t)$ thus yields

$$\begin{aligned}\varepsilon\partial_t\varphi - \varepsilon\Delta\varphi - \tfrac{1}{\varepsilon}\varphi - \tfrac{c_0}{2}g &= \Gamma_\alpha'\big(\partial_t\delta - \Delta\delta - g(\cdot - \delta\nabla\delta, \cdot)\big)\\ &\quad - \tfrac{1}{\varepsilon}\big(\Gamma_\alpha''|\nabla\delta|^2 + \Gamma_\alpha - \varepsilon g(\Gamma_\alpha' - \tfrac{c_0}{2})\big) + 2\varepsilon G\sigma(t)\Gamma_\alpha'\\ &\quad + \varepsilon^2\eta_\alpha\big(\partial_t g_\alpha - \Delta g_\alpha\big) - 2\varepsilon\eta_\alpha'\nabla g_\alpha\cdot\nabla\delta + \Gamma_\alpha'\big(g(\cdot - \delta\nabla\delta, \cdot) - g\big).\end{aligned} \tag{5.8}$$

We now intend to show, upon suitably selecting $\Lambda > 0$, that the right-hand side of (5.8) is almost nonnegative at $(\mathbf{x}_\alpha, t_\alpha)$. With the aid of (5.4) and (5.7), we readily get

$$\Gamma_\alpha'\big(\partial_t\delta - \Delta\delta - g(\cdot - \delta\nabla\delta, \cdot)\big) \geq 0 \qquad \text{at } (\mathbf{x}_\alpha, t_\alpha).$$

Since $(\mathbf{x}_\alpha, t_\alpha)$ is a minimum for $d-\delta$, we have $|\nabla\delta(\mathbf{x}_\alpha, t_\alpha)| = 1$ [9, p. 1104]. In light of (5.6), we distinguish two cases according to whether or not $-\frac{\pi}{2} - \beta < y_\alpha < -\frac{\pi}{2} + \beta$.

4. Let $-\frac{\pi}{2} + \beta < y_\alpha < \frac{\pi}{2} - \beta$. We invoke (3.4) along with (2.2) and (5.3) to arrive at

$$\begin{aligned}&\Gamma_\alpha''(y_\alpha; \mathbf{x}_\alpha, t_\alpha) + \Gamma_\alpha(y_\alpha; \mathbf{x}_\alpha, t_\alpha) - \varepsilon g(\mathbf{x}_\alpha, t_\alpha)\big(\Gamma_\alpha'(y_\alpha; \mathbf{x}_\alpha, t_\alpha) - \tfrac{c_0}{2}\big)\\ &\quad = \int_{-\pi/2}^{\pi/2}\Big(\Gamma''(s; \mathbf{x}_\alpha, t_\alpha) + \Gamma(s; \mathbf{x}_\alpha, t_\alpha) - \varepsilon g(\mathbf{x}_\alpha, t_\alpha)\big(\Gamma'(s; \mathbf{x}_\alpha, t_\alpha) - \tfrac{c_0}{2}\big)\Big)\\ &\qquad \times \zeta_\alpha(y_\alpha - s)ds + \mathcal{O}(\alpha) = \gamma_\alpha'\mathcal{O}(\varepsilon^2) + \mathcal{O}(\alpha).\end{aligned}$$

Hereafter, functions are evaluated at $y_\alpha, \mathbf{x}_\alpha$, and t_α. Because of (2.2) we get

$$\varepsilon^2\eta_\alpha\big(\partial_t g_\alpha - \Delta g_\alpha\big) - 2\varepsilon\eta_\alpha'\nabla g_\alpha\cdot\nabla\delta = \gamma_\alpha'\mathcal{O}(\varepsilon).$$

Since $|y_\alpha| = \big|\delta(\mathbf{x}_\alpha, t_\alpha)/\varepsilon - \frac{\pi}{2} - \sigma(t_\alpha)\big| < \frac{\pi}{2}$, we have $0 < \delta(\mathbf{x}_\alpha, t_\alpha) < \varepsilon(\sigma(t_\alpha) + \pi)$ and

$$\Gamma_\alpha'\big(g(\cdot - \delta\nabla\delta, \cdot) - g\big) \leq G\delta\Gamma_\alpha' \leq \varepsilon G\Gamma_\alpha'\big(\sigma(t_\alpha) + \pi\big).$$

Therefore, in view of (5.4), we can choose Λ sufficiently large in such a fashion that (5.8) becomes

$$\varepsilon\partial_t\varphi - \varepsilon\Delta\varphi - \tfrac{1}{\varepsilon}\varphi - \tfrac{c_0}{2}g \geq \varepsilon\Gamma_\alpha'\big(G\Lambda + \mathcal{O}(1)\big) + \mathcal{O}(\tfrac{\alpha}{\varepsilon}) \geq -C\tfrac{\alpha}{\varepsilon} \qquad \text{at } (\mathbf{x}_\alpha, t_\alpha). \tag{5.9}$$

5. Let $-\frac{\pi}{2} - \beta < y_\alpha \leq -\frac{\pi}{2} + \beta$. The fact $\Gamma(\cdot; \mathbf{x}, t) \in C^{1,1}(\mathbb{R})$ can be used to write the above integral representation of the second term in (5.8) again, but now it is computed in $(-\frac{\pi}{2} - 2\beta, -\frac{\pi}{2} + 2\beta)$. We easily see that such an integral is of order $\mathcal{O}(\beta)$. Choosing $\Lambda = \mathcal{O}(1)$ appropriately, we again conclude (5.9) but with β instead of α.

Assertion (5.1) is just a consequence of $\varphi \in C^\infty(Q), (\mathbf{x}_\alpha, t_\alpha) \to (\mathbf{x}_0, t_0)$ as $\alpha \leq \beta \downarrow 0$, and (5.9). We finally summarize the preceding derivation as follows. Note that we can construct a subsolution along the same lines.

THEOREM 5.1. *The function u_ε^+ defined in* (4.1) *is a viscosity supersolution of the double obstacle problem* (2.7).

6. Comparison. The following result is a crucial tool in comparing viscosity barriers of (2.7), which in particular implies uniqueness for (2.7).

LEMMA 6.1 (comparison lemma). *Let u_ε^+ be a lower semicontinuous viscosity supersolution and u_ε^- be an upper semicontinuous viscosity subsolution. If $u_\varepsilon^+ \geq u_\varepsilon^-$ at $t = 0$ and on $\partial\Omega \times (0,T)$, then $u_\varepsilon^+ \geq u_\varepsilon^-$ for all $(\mathbf{x},t) \in Q$.*

Proof. We split the proof into several steps.

1. Let $\hat{u}_\varepsilon^\pm := \exp(-\lambda t) u_\varepsilon^\pm$, $\hat{g} := \exp(-\lambda t) g$, and $\hat{\lambda} := \varepsilon\lambda - \frac{1}{\varepsilon} > 0$ for $\lambda > 0$ to be chosen later. Then $\hat{u}_\varepsilon^+$ satisfies

$$\varepsilon \partial_t \hat{u}_\varepsilon^+ - \varepsilon \Delta \hat{u}_\varepsilon^+ + \hat{\lambda} \hat{u}_\varepsilon^+ - \tfrac{c_0}{2} \hat{g} \geq 0 \qquad \text{in } \{\hat{u}_\varepsilon^+ < \exp(-\lambda t)\}$$

in the viscosity sense [9, p. 1116], and so does $\hat{u}_\varepsilon^-$, but with reversed inequalities. Since no confusion is possible, we set $u^\pm = \hat{u}_\varepsilon^\pm$ and $g = \hat{g}$.

2. We argue by contradiction. Suppose $u^+ < u^-$ somewhere and set

$$\nu := \sup_{(\mathbf{x},t)\in Q} (u^- - u^+) > 0.$$

Since $u^- - u^+$ is upper semicontinuous and Q is bounded, such a sup is attained in $\overline{Q}$, say, at $(\mathbf{x}_0, t_0)$. The fact that $u^- - u^+ \leq 0$ on $\partial\Omega \times (0,T)$ and $\Omega \times \{0\}$ implies $\mathbf{x}_0 \in \Omega$ and $t_0 > 0$. We do not know whether u^- and u^+ are smooth at $(\mathbf{x}_0, t_0)$ (not even the solution u_ε of (2.7), because u_ε is not better than $W_\infty^{2,1}(Q)$ [11]). We use then a standard argument in the theory of viscosity solutions [8], namely, we double the number of variables and at the same time penalize the doubling. Consider the upper semicontinuous function

$$U(\mathbf{x},\mathbf{y},t,s) := u^-(\mathbf{x},t) - u^+(\mathbf{y},s) - \tfrac{1}{\alpha}\big(|\mathbf{x}-\mathbf{y}|^2 + |t-s|^2\big)$$

for $\alpha \downarrow 0$. Since $U(\mathbf{x},\mathbf{x},t,t) = u^-(\mathbf{x},t) - u^+(\mathbf{x},t)$, we see that

$$2 \geq \nu_\alpha := \sup_{(\mathbf{x},\mathbf{y},t,s)\in Q^2} U(\mathbf{x},\mathbf{y},t,s) \geq \nu > 0.$$

Since Q is bounded, we conclude that there exists a point $(\mathbf{x}_\alpha, \mathbf{y}_\alpha, t_\alpha, s_\alpha) \in \overline{Q}^2$ at which the sup is attained, $U(\mathbf{x}_\alpha, \mathbf{y}_\alpha, t_\alpha, s_\alpha) = \nu_\alpha$. An application of Lemma 3.1 in [8, p. 15] yields

$$\tfrac{1}{\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}_\alpha|^2 + |t_\alpha - s_\alpha|^2\big) \to_{\alpha\downarrow 0} 0, \tag{6.1}$$

$$\nu_\alpha \downarrow_{\alpha\downarrow 0} \nu = (u^- - u^+)(\mathbf{x}_0, t_0). \tag{6.2}$$

Here we report the argument for completeness. As $\alpha \downarrow 0$ we clearly have that $\{\nu_\alpha\}$ is decreasing, which in turn shows the existence of $\lim_{\alpha\downarrow 0} \nu_\alpha$. Since

$$\begin{aligned} \nu_{2\alpha} &\geq u^-(\mathbf{x}_\alpha, t_\alpha) - u^+(\mathbf{y}_\alpha, s_\alpha) - \tfrac{1}{2\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}_\alpha|^2 + |t_\alpha - s_\alpha|^2\big) \\ &= \nu_\alpha + \tfrac{1}{2\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}_\alpha|^2 + |t_\alpha - s_\alpha|^2\big), \end{aligned}$$

we get

$$\tfrac{1}{\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}_\alpha|^2 + |t_\alpha - s_\alpha|^2\big) \leq 2(\nu_{2\alpha} - \nu_\alpha) \to_{\alpha\downarrow 0} 0.$$

This proves (6.1). Since $(\mathbf{x}_\alpha, \mathbf{y}_\alpha, t_\alpha, s_\alpha)$ belongs to a compact set, we can extract a subsequence, still labeled as before, such that

$$(\mathbf{x}_\alpha, \mathbf{y}_\alpha, t_\alpha, s_\alpha) \to_{\alpha\downarrow 0} (\overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{t}, \overline{s}) = (\overline{\mathbf{x}}, \overline{\mathbf{x}}, \overline{t}, \overline{t}).$$

Since $u^- - u^+$ is upper semicontinuous we deduce

$$\nu \geq u^-(\bar{\mathbf{x}}, \bar{t}) - u^+(\bar{\mathbf{x}}, \bar{t}) \geq \limsup_{\alpha \downarrow 0} U(\mathbf{x}_\alpha, \mathbf{y}_\alpha, t_\alpha, s_\alpha) = \lim_{\alpha \downarrow 0} \nu_\alpha \geq \nu.$$

So (6.2) follows and $(\bar{\mathbf{x}}, \bar{t})$ can be identified with $(\mathbf{x}_0, t_0)$ above; thus $\bar{\mathbf{x}} \in \Omega, \bar{t} > 0$. Moreover, from (6.1) we see that

$$\mathbf{x}_\alpha, \mathbf{y}_\alpha \in \Omega, \qquad t_\alpha, s_\alpha > 0, \tag{6.3}$$

$$u^-(\mathbf{x}_\alpha, t_\alpha) > -\exp(-\lambda t_\alpha), \qquad u^+(\mathbf{y}_\alpha, s_\alpha) < \exp(-\lambda s_\alpha). \tag{6.4}$$

3. Set

$$U^+(\mathbf{x}, t) := u^+(\mathbf{y}_\alpha, s_\alpha) + \tfrac{1}{\alpha}\big(|\mathbf{x} - \mathbf{y}_\alpha|^2 + |t - s_\alpha|^2\big) + \nu_\alpha$$

and note that $U(\mathbf{x}, \mathbf{y}_\alpha, t, s_\alpha) = u^-(\mathbf{x}, t) - U^+(\mathbf{x}, t) + \nu_\alpha$. We then realize that $u^- - U^+$ attains a maximum at $(\mathbf{x}_\alpha, t_\alpha)$, and from (2.8), (6.3), and (6.4) we infer that

$$\varepsilon \partial_t U^+ - \varepsilon \Delta_{\mathbf{x}} U^+ + \hat{\lambda} U^+ - \tfrac{c_0}{2} g \leq 0 \qquad \text{at } (\mathbf{x}_\alpha, t_\alpha);$$

thus

$$\tfrac{2\varepsilon}{\alpha}(t_\alpha - s_\alpha) - \tfrac{2n\varepsilon}{\alpha} + \hat{\lambda}\Big(u^+(\mathbf{y}_\alpha, s_\alpha) + \tfrac{1}{\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}_\alpha|^2 + |t_\alpha - s_\alpha|^2\big) + \nu_\alpha\Big) \leq \tfrac{c_0}{2} g(\mathbf{x}_\alpha, t_\alpha). \tag{6.5}$$

4. Set

$$U^-(\mathbf{y}, s) := u^-(\mathbf{x}_\alpha, t_\alpha) - \tfrac{1}{\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}|^2 + |t_\alpha - s|^2\big) - \nu_\alpha$$

and note that $U(\mathbf{x}_\alpha, \mathbf{y}, t_\alpha, s) = U^-(\mathbf{y}, s) - u^+(\mathbf{y}, s) + \nu_\alpha$. The function $u^+ - U^-$ attains a minimum at $(\mathbf{y}_\alpha, s_\alpha)$, which in view of (2.8), (6.3), and (6.4) yields

$$\varepsilon \partial_s U^- - \varepsilon \Delta_{\mathbf{y}} U^- + \hat{\lambda} U^- - \tfrac{c_0}{2} g \geq 0 \qquad \text{at } (\mathbf{y}_\alpha, s_\alpha).$$

This can be written as

$$\tfrac{2\varepsilon}{\alpha}(t_\alpha - s_\alpha) + \tfrac{2n\varepsilon}{\alpha} + \hat{\lambda}\Big(u^-(\mathbf{x}_\alpha, t_\alpha) - \tfrac{1}{\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}_\alpha|^2 + |t_\alpha - s_\alpha|^2\big) - \nu_\alpha\Big) \geq \tfrac{c_0}{2} g(\mathbf{y}_\alpha, s_\alpha). \tag{6.6}$$

5. Now subtract (6.5) from (6.6) to get

$$\begin{aligned} \tfrac{4n\varepsilon}{\alpha} + \hat{\lambda}\Big(U(\mathbf{x}_\alpha, \mathbf{y}_\alpha, t_\alpha, s_\alpha) - \tfrac{1}{\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}_\alpha|^2 + |t_\alpha - s_\alpha|^2\big) - 2\nu_\alpha\Big) \\ + \tfrac{c_0}{2}\big(g(\mathbf{x}_\alpha, t_\alpha) - g(\mathbf{y}_\alpha, s_\alpha)\big) \geq 0. \end{aligned} \tag{6.7}$$

In view of (6.1) and (6.2), we can choose α so small that

$$U(\mathbf{x}_\alpha, \mathbf{y}_\alpha, t_\alpha, s_\alpha) - \tfrac{1}{\alpha}\big(|\mathbf{x}_\alpha - \mathbf{y}_\alpha|^2 + |t_\alpha - s_\alpha|^2\big) - 2\nu_\alpha = -\nu_\alpha + o(1) < -\tfrac{1}{2}\nu$$

and $\tfrac{c_0}{2}\big|g(\mathbf{x}_\alpha, t_\alpha) - g(\mathbf{y}_\alpha, s_\alpha)\big| \leq \tfrac{4n\varepsilon}{\alpha}$. Finally, select $\lambda > \tfrac{16n}{\alpha\nu} + 1/\varepsilon^2$, which yields $\tfrac{1}{2}\hat{\lambda}\nu = \tfrac{1}{2}\big(\varepsilon\lambda - \tfrac{1}{\varepsilon}\big)\nu > \tfrac{8n\varepsilon}{\alpha}$. Then (6.7) is written as

$$0 < \tfrac{8n\varepsilon}{\alpha} - \tfrac{1}{2}\big(\varepsilon\lambda - \tfrac{1}{\varepsilon}\big)\nu < 0,$$

which is a contradiction. This proves the assertion $u^+ \geq u^-$ in Q. []

7. Convergence and error estimates. Our goal is to prove convergence of the zero-level sets $\Sigma_\varepsilon(t)$ to the generalized geometric motion $\Sigma(t)$. Key ingredients are the explicit form of supersolutions of §5 and the comparison lemma of §6.

Let $l_\varepsilon := (\Lambda e^{2GT} + \pi)\varepsilon = (\sigma(0) + \pi)\varepsilon$ and $\Sigma^\pm_{0,\varepsilon} := \{\mathbf{x} \in \Omega : d_0(\mathbf{x}) = \mp l_\varepsilon\}$, which is of class C^1. We designate by $\Sigma^\pm_\varepsilon(t)$ the generalized evolving fronts that, starting from $\Sigma^\pm_{0,\varepsilon}$, are governed by (1.2), namely,

$$\Sigma^\pm_\varepsilon(t) = \{\mathbf{x} \in \Omega : \omega(\mathbf{x}, t) = \mp l_\varepsilon\},$$

where ω is the (unique) continuous viscosity solution of (2.3). Let $d^\pm_\varepsilon(\cdot, t)$ be the corresponding signed distance functions to $\Sigma^\pm_\varepsilon(t)$ and $u^\pm_\varepsilon$ be the barriers just constructed in §4 in terms of $d^\pm_\varepsilon$ (see (4.1) and (4.5)). If $\Sigma^\pm_\varepsilon(t)$ develops interior, then $d^+_\varepsilon(\cdot, t)$ must be replaced by $\operatorname{dist}(\cdot, \{\omega(\cdot, t) < -l_\varepsilon\})$ and $d^-_\varepsilon(\cdot, t)$ by $-\operatorname{dist}(\cdot, \{\omega(\cdot, t) > l_\varepsilon\})$ for the argument below to apply (see (2.4)). We would like to prove that

$$u^-_\varepsilon(\mathbf{x}, t) \le u_\varepsilon(\mathbf{x}, t) \le u^+_\varepsilon(\mathbf{x}, t) \qquad \forall\ (\mathbf{x}, t) \in Q. \tag{7.1}$$

In light of Lemma 6.1 and the fact that u^+_ε is lower semicontinuous and u^-_ε is upper semicontinuous, because of (2.6), we only have to show that (7.1) is valid on the parabolic boundary of Q. This is certainly the case on $\partial\Omega \times (0, T)$, simply because $u^-_\varepsilon = u^+_\varepsilon = u_\varepsilon = 1$ in this set. In addition, since $d^\pm_\varepsilon(\mathbf{x}, 0) = d_0(\mathbf{x}) \pm l_\varepsilon$, for $t = 0$ and $\mathbf{x} \in \Omega$ we have

$$u^+_\varepsilon(\mathbf{x}, 0) = \Gamma\big(\tfrac{d^+_\varepsilon(\mathbf{x},0)}{\varepsilon} - \tfrac{\pi}{2} - \sigma(0)\big) = \Gamma\big(\tfrac{d_0(\mathbf{x})}{\varepsilon} + \tfrac{\pi}{2}\big) \ge \operatorname{sign}\big(d_0(\mathbf{x})\big) = u_\varepsilon(\mathbf{x}, 0) \ge u^-_\varepsilon(\mathbf{x}, 0).$$

The desired inequality (7.1) then follows immediately from Lemma 6.1.

THEOREM 7.1. *For $\mathbf{x} \in I(t)$ (resp., $\mathbf{x} \in O(t)$), there exists $\varepsilon_0(\mathbf{x}, t) > 0$ such that*

$$u_\varepsilon(\mathbf{x}, t) = -1 \qquad (\textit{resp.}, \ u_\varepsilon(\mathbf{x}, t) = +1) \qquad \forall\ \varepsilon \le \varepsilon_0(\mathbf{x}, t).$$

Proof. Let $\mathbf{x} \in I(t) = \{\omega(\cdot, t) < 0\}$. For ε sufficiently small, $\omega(\mathbf{x}, t) < -l_\varepsilon$, whence $d^+_\varepsilon(\mathbf{x}, t) < 0$. This implies $d^+_\varepsilon(\mathbf{x}, t)/\varepsilon - \frac{\pi}{2} - \sigma(t) < -\frac{\pi}{2}$ and, therefore, $-1 \le u_\varepsilon(\mathbf{x}, t) \le u^+_\varepsilon(\mathbf{x}, t) = -1$, because of (7.1). Similar reasoning for u^-_ε completes the proof. □

Remark 7.1. If $\Sigma(t)$ has an empty interior, then Theorem 7.1 establishes the convergence of $\Sigma_\varepsilon(t)$ to $\Sigma(t)$. As far as we know, the question of whether such a condition is always valid for the evolution of smooth initial surfaces by mean curvature, namely, $g = 0$, remains a conjecture [10]. Instead, if $g \ne 0$, then $\Sigma(t)$ may develop interior [4].

To derive interface error estimates we are forced to assume more regularity of $\Sigma(t)$. We say that $\mathbf{x} \in \Sigma^*(t)$, the *regular* part of $\Sigma(t)$, if $\mathbf{x} \in \Sigma(t)$ and $\omega(\cdot, t)$ is of class C^1 in a neighborhood of $\mathbf{x}$ and satisfies the nondegeneracy condition $|\nabla\omega(\mathbf{x}, t)| > 0$. Note that ω is only known to be Lipschitz continuous in Q and also that $\Sigma(t) = \Sigma^*(t)$ as long as the motion is classical, that is, before the onset of singularities. Nevertheless, $\Sigma(t) = \Sigma^*(t)$ is known to hold between consecutive singularities for a number of flows, e.g., for surfaces of rotation [2], [18].

Let $\operatorname{thick}\big(\mathcal{T}_\varepsilon(t); \mathbf{x}, \mathbf{n}\big)$ denote the thickness of the transition region $\mathcal{T}_\varepsilon(t) := \{\mathbf{x} \in \Omega : |u_\varepsilon(\mathbf{x}, t)| < 1\}$ in the normal direction $\mathbf{n} := \nabla\omega(\mathbf{x}, t)/|\nabla\omega(\mathbf{x}, t)|$ across $\mathbf{x} \in \Sigma(t)$.

THEOREM 7.2. *For $\mathbf{x} \in \Sigma^*(t)$, there exists $\varepsilon_0(\mathbf{x}, t) > 0$ such that*

$$\operatorname{dist}\big(\mathbf{x}, \Sigma_\varepsilon(t)\big), \tfrac{1}{2}\operatorname{thick}\big(\mathcal{T}_\varepsilon(t); \mathbf{x}, \mathbf{n}\big) \le 2(\Lambda e^{2GT} + \pi)|\nabla\omega(\mathbf{x}, t)|^{-1}\varepsilon \quad \forall\ \varepsilon \le \varepsilon_0(\mathbf{x}, t). \tag{7.2}$$

Proof. Since $u^\pm_\varepsilon(\cdot, t) = \mp 1$ on $\Sigma^\pm_\varepsilon(t)$ and (7.1) is valid, it suffices for us to estimate the distance between $\Sigma^\pm_\varepsilon(t)$ and $\mathbf{x} \in \Sigma^*(t)$. Using Taylor's formula about $\mathbf{x}$, for $\mathbf{x} + \delta\mathbf{n} \in \Sigma^-_\varepsilon(t)$, and so $\delta > 0$, and ε sufficiently small depending on $(\mathbf{x}, t)$, we obtain

$$l_\varepsilon = \omega\big(\mathbf{x} + \delta\mathbf{n}, t\big) = \omega(\mathbf{x}, t) + |\nabla\omega(\mathbf{x}, t)|\delta + o(\delta) \ge \tfrac{1}{2}|\nabla\omega(\mathbf{x}, t)|\delta.$$

Hence $\delta \le 2l_\varepsilon |\nabla\omega(\mathbf{x},t)|^{-1}$, as asserted in (7.2). □

Remark 7.2. The exponential blowup of the constant in (7.2) can only be avoided if g is independent of $\mathbf{x}$ (see Remark 4.2). In fact, (7.2) cannot be improved without additional assumptions, as the following radially symmetric flow in two dimensions reveals. Let $g(r,t) = r$ and consider the initial condition $r_\delta(0) = 1 + \delta$. The corresponding evolution is given by $r_\delta(t) = \big(1 + (r_\delta^2(0) - 1)e^{2t}\big)^{1/2}$, which yields $r_\delta(t) - r_0(t) = \mathcal{O}\big(\delta e^{2t}\big)$.

Remark 7.3. Chen and Elliott [6] derived a linear rate of convergence for interfaces before the onset of singularities, say in $[0,t^*)$, that is valid for the mean curvature flow ($g = 0$):

$$\operatorname{dist}_H\big(\Sigma(t),\Sigma_\varepsilon(t)\big) \le C_{t^\#}\varepsilon \qquad \forall\, t \le t^\# < t^*;$$

here dist_H stands for the Hausdorff distance. In such a regime, however, optimal error estimates of order $\mathcal{O}(\varepsilon^2)$ were established in [14], [15]. The virtue of (7.2) is thus its validity even beyond singularities if the motion is locally smooth.

Remark 7.4. We stress that perturbing the original motion (1.2) as in [3], that is, replacing g by $g \pm o(1)$, would not immediately lead to (7.2). This is true because $d_\varepsilon^\pm$ would depend on the viscosity solutions $\omega_\varepsilon^\pm$ of the resulting perturbed problems rather than on ω. Even though convergence of $\omega_\varepsilon^\pm$ to ω is known as $\varepsilon \downarrow 0$, we would also need uniform nondegeneracy of $\omega_\varepsilon^\pm$ to hold, which does not seem to be available.

Remark 7.5. We can summarize Theorem 7.2 by simply saying that $\Sigma_\varepsilon(t)$ lies between the surfaces $\omega(\cdot,t) = \pm l_\varepsilon = \pm\mathcal{O}(\varepsilon)$. It is then the profile of ω near a singularity or, equivalently, how far the level sets $\Sigma_\varepsilon^\pm(t)$ may separate from each other in the vicinity of a singular point, that determines the rate of convergence.

REFERENCES

[1] S. M. Allen and J.W. Cahn, *A macroscopic theory for antiphase boundary motion and its application to antiphase domain coarsing*, Acta Metall., 27 (1979), pp. 1085–1095.

[2] S. Altschuler, S. Angenent, and Y. Giga, *Mean curvature flow through singularities for surfaces of rotation*, J. Geom. Anal., to appear.

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

[4] G. Bellettini and M. Paolini, *Two examples of fattening for the curvature flow with a driving force*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. (9) Mat. Appl., 5 (1994), pp. 229–236.

[5] G. Caginalp, *An analisys of a phase field model of a free boundary*, Arch. Rational Mech. Anal., 92 (1986), pp. 205–245.

[6] X. Chen and C.M. Elliott, *Asymptotics for a parabolic double obstacle problem*, Proc. Roy. Soc. London Ser. A, 444 (1994), pp. 429–445.

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

[8] M.G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[9] L.C. Evans, H.-M. Soner, and P.E. Souganidis, *Phase transitions and generalized motion by mean curvature*, Comm. Pure Appl. Math., 45 (1992), pp. 1097–1123.

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

[11] A. Friedman, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[12] Y. Giga, S. Goto, H. Ishii, and M.H. Sato, *Comparison principle and convexity preserving properties for singular degenerate parabolic equations on unbounded domains*, Indiana

Univ. Math. J., 40 (1991), pp. 443–470.

[13] R.H. NOCHETTO, M. PAOLINI, S. ROVIDA, AND C. VERDI, *Variational approximation of the geometric motion of fronts*, in Motion by Mean Curvature and Related Topics, G. Buttazzo and A. Visintin, eds., Gruyter, Berlin, 1994, pp. 124–149.

[14] R.H. NOCHETTO, M. PAOLINI, AND C. VERDI, *Optimal interface error estimates for the mean curvature flow*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 21 (1994), pp. 193–212.

[15] ———, *Sharp error analysis for curvature dependent evolving fronts*, Math. Models Methods Appl. Sci., 3 (1993), pp. 771–723.

[16] ———, *Double obstacle formulation with variable relaxation parameter for smooth geometric front evolutions: Asymptotic interface error estimates*, Asymptotic Anal., 10 (1995), pp. 173–198.

[17] M. PAOLINI AND C. VERDI, *Asymptotic and numerical analyses of the mean curvature flow with a space-dependent relaxation parameter*, Asymptotic Anal., 5 (1992), pp. 553–574.

[18] H.-M. SONER AND P.E. SOUGANIDIS, *Singularities and uniqueness of cylindrically symmetric surfaces moving by mean curvature*, Comm. Partial Differential Equations, 18 (1993), pp. 859–894.

[19] N. YAMADA, *Viscosity solutions for a system of elliptic inequalities with bilateral obstacles*, Funkcial. Ekvac., 30 (1987), pp. 417–425.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1527–1546, November 1995

009

INSTABILITY AND BLOW-UP OF SOLUTIONS TO A GENERALIZED BOUSSINESQ EQUATION*

YUE LIU†

Abstract. In this paper we investigate conditions for the finite-time blow-up of solutions of the generalized Boussinesq equation (BQ)

$$u_{tt} - u_{xx} + (f(u) + u_{xx})_{xx} = 0, \qquad x \in \mathbf{R}, t > 0.$$

The conditions are expressed in terms of the energy of the ground state. In particular, there exist initial data arbitrarily close to the stationary state of lowest energy whose solutions blow up in finite time.

Key words. Boussinesq equation, blow-up, solitary waves, ground state, stability theory

AMS subject classifications. 35Q, 35B, 76B

1. Introduction.

It is well known that the initial-value problem of the Boussinesq equation is not always globally well posed. There exist smooth initial wave and velocity profiles for which the solution that emanates from them loses regularity in a finite time [28], [16]. The purpose of this paper is to investigate some general conditions for the existence and nonexistence of global solutions to a generalized Boussinesq equation (BQ)

$$u_{tt} - u_{xx} + (f(u) + u_{xx})_{xx} = 0, \tag{1}$$

where $f \in C^1(\mathbf{R})$ with $f(0) = 0$. Equation (1) has the equivalent form

$$\begin{cases} u_t = v_x, \\ v_t = (u - u_{xx} - f(u))_x. \end{cases} \tag{2}$$

It has the four natural invariants

$$E(\vec{u}) = E(u,v) = \int_{-\infty}^{\infty} \left(\frac{1}{2}\, u^2 + \frac{1}{2}\, u_x^2 + \frac{1}{2}\, v^2 - F(u) \right) dx,$$

$$V(\vec{u}) = V(u,v) = \int_{-\infty}^{\infty} uv\, dx,$$

$$I_1(\vec{u}) = I_1(u,v) = \int_{-\infty}^{\infty} u\, dx,$$

$$I_2(\vec{u}) = I_2(u,v) = \int_{-\infty}^{\infty} v\, dx,$$

where we write $\vec{u}$ to denote the pair (u, v) and where $F' = f$ and $F(0) = 0$.

From the general methods of Levine [34] used by Kalantarov and Ladyzhenskaya in [28], a solution of the BQ with nonpositive-energy initial data blows up in some

* Received by the editors November 1, 1993; accepted for publication (in revised form) April 8, 1994. This research was supported in part by National Science Foundation grant DMS 90-23864 and Army Research Office grant DAAH 04-93-G0198.

† Department of Mathematics, University of Texas at Austin, Austin, Texas 78712.

sense. In this paper we investigate more general conditions for blow up and, in particular, the relationship to the ground state.

We use $|\cdot|_p$ to denote the norm in $L^p(\mathbf{R}^n)$ and $\|\cdot\|_s$ the norm in Sobolev space $H^s(\mathbf{R}^n)$.

First, in §2, we determine the best constant $C_{p,n}$ for the Sobolev inequality

$$|u|_{p+1} \le C_{p,n}\|u\|_1, \tag{3}$$

for $1 < p < \gamma(n)$, where $\gamma(n) = \frac{n+2}{n-2}$ for $n \ge 3(\gamma(1) = \gamma(2) = \infty)$. We obtain (Corollary 2.5)

$$C_{p,n} = \|\varphi\|_1^{-\frac{p-1}{p+1}}, \tag{4}$$

where φ_n is the positive radial H^1 solution (the ground state) of

$$-\Delta\varphi_n + \varphi_n - |\varphi_n|^{p-1}\varphi_n = 0, \tag{5}$$

for $x \in \mathbf{R}^n$ and $1 < p < \gamma(n)$. We know, when $n = 1, \varphi(x) \equiv \varphi_1(x) = \left(\frac{p+1}{2}\right)^{1/(p-1)} \operatorname{sech}^{2/(p-1)}\left(\frac{p-1}{2}x\right)$ for $1 < p < \infty$. The best constant of an interpolation estimate among various norms often has an analytical or geometrical significance [1], [24] and has been much studied [6], [18], [19], [36], [41], [43], [44], [45], [54], [56]. It is possible that our result is already known, but we could not find it in the literature. The result (4) is derived from the following considerations. To compute $C_{p,n}$, it suffices to minimize the functional

$$I_{p,n}(u) = \frac{\|u\|_1^{p+1}}{|u|_{p+1}^{p+1}} \tag{6}$$

for $1 < p < \gamma(n)$. We show that the minimum is attained at the ground state φ_n. It follows that the minimum d_n of the energy $E(u, 0)$ subject to the constraint that $R(u) = 0$ is attained at the ground state φ_n, where we define

$$R(u) \equiv \|u\|_1^2 - |u|_{p+1}^{p+1}. \tag{7}$$

In §3, we follow the main idea of [46] to construct invariant regions K_1 and K_2 for the flow governed by BQ, where

$$\begin{aligned} K_1 &\equiv \{u \in H^1(\mathbf{R}) | E(u,0) < d, R(u) > 0\}, \\ K_2 &\equiv \{u \in H^1(\mathbf{R}) | E(u,0) < d, R(u) < 0\} \end{aligned} \tag{8}$$

with $d = d_1$. Then we obtain conditions on the initial data for which the BQ has global bounded solution (Theorem 3.4) or for which the solution blows up in a finite time (Theorems 4.1 and 4.3). The energy of the initial data for which a solution blows up does not have to be negative. Furthermore, in Theorem 4.3, the blow-up result implies the instability of the ground state solution.

Finally, in §5, we construct global bounded solutions with initial data in different regions that are related to traveling wave solutions of the BQ.

2. Variational problem and ground state. We study the nonlinear functional $I_{p,n}$ defined by (6) that is naturally associated with the Sobolev inequality (3). By (3), $I_{p,n}$ is defined on $H^1(\mathbf{R}^n)$.

THEOREM 2.1. *If* $1 < p < \gamma(n)$, *then*

$$I_{p,n}(\varphi_n) = \min_{0 \neq u \in H^1(\mathbf{R}^n)} I_{p,n}(u),$$

where φ_n *is the ground state of* (5).

Remark. $I_{p,n}(\varphi_n) = \|\varphi_n\|_1^{p-1}$. In fact, multiplying the equation by φ_n and integrating, we obtain

$$\|\varphi_n\|_1^2 = |\varphi_n|_{p+1}^{p+1}. \tag{9}$$

Therefore $I_{p,n}(\varphi_n) = \|\varphi_n\|_1^{p-1}$.

The proof of Theorem 2.1 follows from a series of lemmas, which are based on estimates of ground state φ_n. Lemma 2.2 was obtained by M. Weinstein [56] using the compactness lemma of W. Strauss [54].

LEMMA 2.2. *Let* $n \geq 2$ *and* $1 < p < \gamma(n)$. *Let*

$$J_{p,n}(u) \equiv \frac{|u|_2^{2+\frac{p-1}{2}(2-n)} |\nabla u|_2^{\frac{p-1}{2}n}}{|u|_{p+1}^{p+1}} \tag{10}$$

and $\alpha_{p,n} \equiv \min_{0 \neq u \in H^1(\mathbf{R}^n)} J_{p,n}(u)$. *Then* $\alpha_{p,n} = 2|\psi|_2^{p-1}/(p+1)$, *where* ψ *is the ground state of*

$$-\frac{p-1}{4} n\Delta\psi + \left(1 + \frac{p-1}{4}(2-n)\right)\psi - \psi^p = 0.$$

For $n = 1$, the following lemma was obtained by Nagy [41].

LEMMA 2.3. *Let* $n = 1$ *and* $1 < p < \infty$. *Then*

$$|u|_{p+1}^{p+1} \leq \left[H\left(\frac{2}{p-1}, \frac{1}{2}\right)\right]^{\frac{p-1}{2}} |u|_2^{\frac{p+3}{2}} |u_x|_2^{\frac{p-1}{2}},$$

for $u \in H^1(\mathbf{R})$, *where*

$$H(a,b) \equiv \frac{(a+b)^{-(a+b)}\Gamma(1+a+b)}{a^{-a}b^{-b}\Gamma(1+a)\Gamma(1+b)}$$

for $a \geq 0, b \geq 0$.

Remark. $H(a,0) = 1, H(a,1) = \left(1 + \frac{1}{a}\right)^{-a}$, and $H(a,b)$ is decreasing in $b > 0$. Hence $e^{-1} < H(a,b) \leq 1$ for $0 < b < 1$ and $a \geq 0$.

LEMMA 2.4. *For* $n \geq 1$,

$$\begin{aligned}\alpha_{p,n} &= J_{p,n}(\varphi_n) \\ &= [n(p+1)]^{\frac{p-1}{4}n} [2(p+1)]^{-\frac{p+1}{2}} [(n+2) - p(n-2)]^{\frac{1}{4}[(n+2)-p(n-2)]} \|\varphi_n\|_1^{p-1},\end{aligned}$$

where φ_n *is the ground state of* (5).

Proof. Let $\psi(x) = \lambda\varphi_n(\mu x)$, where

$$\lambda^{p-1} = \frac{(n+2) - p(n-2)}{4}, \qquad \mu^2 = \frac{(n+2) - p(n-2)}{n(p-1)}.$$

Then φ_n is the ground state of (5). Using identities

$$|\varphi_n|_2^2 = \frac{(n+2)-p(n-2)}{2(p+1)}\|\varphi_n\|_1^2 \tag{11}$$

and

$$|\nabla\varphi_n|_2^2 = \frac{n(p-1)}{2(p+1)}\|\varphi_n\|_1^2, \tag{12}$$

it is easy to compute

$$\begin{aligned} J_{p,n}(\varphi_n) = \|\varphi_n\|_1^{2+\frac{p-1}{2}(2-n)} &\left[\frac{(n+2)-p(n-2)}{2(p+1)}\right]^{\frac{1}{4}[(n+2)-p(n-2)]} \\ &\cdot\left[\frac{n(p-1)}{2(p+1)}\right]^{\frac{p-1}{4}n}\|\varphi_n\|_1^{\frac{p-1}{2}n-2} \quad \text{for } n \geq 1 \end{aligned} \tag{13}$$

and obtain from Lemma 2.2 that $J_{p,n}(\varphi) = \alpha_{p,n}$ for $n \geq 2$. For the case of $n = 1$ and $1 < p < \infty$,

$$J_{p,1}^{-\frac{2}{p-1}}(\varphi) = H\left(\frac{2}{p-1}, \frac{1}{2}\right), \tag{14}$$

where $\varphi(x) = \left(\frac{p+1}{2}\right)^{\frac{1}{(p-1)}} \operatorname{sech}^{\frac{2}{(p-1)}}\left(\frac{p-1}{2}x\right)$. In fact, putting $n = 1$ in (13) and using identity (11), we have

$$J_{p,1}^{-\frac{2}{p-1}}(\varphi) = 2^{-\frac{p-5}{p-1}}(p+3)^{\frac{p-5}{2(p-1)}}(p-1)^{\frac{1}{2}}\left[\int_{-\infty}^{\infty}\operatorname{sech}^{\frac{4}{p-1}}(x)\,dx\right]^{-1}. \tag{15}$$

On the other hand, it is easy to compute

$$\begin{aligned} H\left(\frac{2}{p-1}, \frac{1}{2}\right) &\equiv \frac{\left(\frac{2}{p-1}+\frac{1}{2}\right)^{-\left(\frac{2}{p-1}+\frac{1}{2}\right)}\Gamma\left(1+\frac{2}{p-1}+\frac{1}{2}\right)}{\left(\frac{2}{p-1}\right)\left(\frac{1}{2}\right)^{-\frac{1}{2}}\Gamma\left(1+\frac{2}{p-1}\right)\Gamma\left(\frac{1}{2}+1\right)} \\ &= 2^{\frac{5-p}{p-1}}(p+3)^{\frac{p-5}{2(p-1)}}(p-1)^{\frac{1}{2}}\left[B\left(\frac{1}{2}, \frac{2}{p-1}\right)\right]^{-1}, \end{aligned} \tag{16}$$

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. But

$$\int_{-\infty}^{\infty}\operatorname{sech}^{\frac{4}{p-1}}(x)\,dx = B\left(\frac{1}{2}, \frac{2}{p-1}\right). \tag{17}$$

Comparing this identity with (15) and (16), we obtain (14). Using identity (14), we are then able to prove $\alpha_{p,1} = J_{p,1}(\varphi)$. First, we have

$$\alpha_{p,1} \equiv \min_{0\neq u\in H^1(\mathbf{R})} J_{p,1}(u) \leq J_{p,1}(\varphi).$$

On the other hand, using Lemma 2.2, we have

$$J_{p,1}(u) \geq \left[H\left(\frac{2}{p-1}, \frac{1}{2}\right)\right]^{-\frac{p-1}{2}} \quad \text{for all } u \in H^1(\mathbf{R}),$$

and using the identity (14), we obtain

$$\alpha_{p,1} \geq \left[H\left(\frac{2}{p-1}, \frac{1}{2}\right)\right]^{-\frac{p-1}{2}} = J_{p,1}(\varphi).$$

Hence, $\alpha_{p,1} = J_{p,1}(\varphi)$. Combining this equality with the proof of the case $n \geq 2$, the proof of Lemma 2.4 is completed. □

Using Lemma 2.4 and the idea in [36], we are now able to prove Theorem 2.1.

Proof of Theorem 2.1. If $u_\lambda = u(\lambda x)$, for $u \in H^1(\mathbf{R}^n)$, then

$$[I_{p,n}(u_\lambda)]^{\frac{2}{p+1}} = \frac{\lambda^{-n}|u|_2^2 + \lambda^{2-n}|\nabla u|_2^2}{\lambda^{-\frac{2n}{p+1}}|u|_{p+1}^2} = \lambda^\alpha A + \lambda^{\alpha+2}B,$$

where

$$A = \frac{|u|_2^2}{|u|_{p+1}^2}, \qquad B = \frac{|\nabla u|_2^2}{|u|_{p+1}^2},$$

$$\alpha = \frac{1-p}{1+p}\, n < 0, \quad \text{and} \quad \alpha + 2 = \frac{1}{p+1}\,[n + 2 - p(n-2)] > 0.$$

Let $F(\alpha) = \lambda^\alpha A + \lambda^{\alpha+2}B$ for $0 < \lambda < \infty$. We shall optimize λ. It is easy to compute

$$F'(\lambda) = \alpha\lambda^{\alpha-1}A + (\alpha+2)\lambda^{\alpha+1}B,$$
$$F''(\lambda) = \alpha(\alpha-1)\lambda^{\alpha-2}A + (\alpha+2)(\alpha+1)\lambda^\alpha B.$$

Solving $F'(\lambda) = 0$ yields

$$\lambda = \lambda_0 = \left(-\frac{\alpha A}{(\alpha+2)B}\right)^{\frac{1}{2}} > 0$$

and

$$F''(\lambda_0) = \lambda_0^{\alpha-2}\left[\alpha(\alpha-1)A + (\alpha+1)(\alpha+2)\left(-\frac{\alpha A}{(\alpha+2)B}\right)B\right]$$
$$= \lambda_0^{\alpha-2}(-2\alpha)A > 0.$$

Since $F(\lambda) \to +\infty$ as $\lambda \to \phi$ and $\lambda \to +\infty$, we obtain

$$\min_{0<\lambda<\infty} F(\lambda) = F(\lambda_0) = \frac{2}{\alpha+2}\,\lambda_0^\alpha A.$$

Thus

$$[I_{p,n}(u_\lambda)]^{\frac{2}{p+1}} \equiv F(\lambda) \geq F(\lambda_0) = \frac{2}{\alpha+2}\lambda_0^\alpha A$$

for $0 < \lambda < +\infty$. Hence

$$I_{p,n}(u) = I_{p,n}(u_1) \geq [F(\lambda_0)]^{\frac{p+1}{2}} = \left(\frac{2}{\alpha+2}\right)^{\frac{p+1}{2}} (\lambda_0^2)^{\frac{\alpha(p+1)}{4}} A^{\frac{p+1}{2}}$$
$$= k(n,p)h(n,p)A^{\frac{1}{4}[(n+2)-p(n-2)]}B^{\frac{p+1}{2}}$$
$$= k(n,p)h(n,p)|u|_{2+\frac{p-1}{2}(2-n)}|\nabla u|_2^{\frac{p-1}{2}n}|u|_{p+1}^{-(p+1)}$$
$$= k(n,p)h(n,p)J_{p,n}(u),$$

where

$$k(n,p)=\left[\frac{2(p+1)}{(n+2)-p(n-2)}\right]^{\frac{p+1}{2}} \quad \text{and} \quad h(n,p)=\left[\frac{(p-1)n}{(n+2)-p(n-2)}\right]^{\frac{1-p}{4}n}.$$

Lemma 2.4 yields

$$\begin{aligned}\min_{0\neq u\in H^1(\mathbf{R}^n)} I_{p,n}(u) &\geq k(n,p)h(n,p)\alpha_{p,n}\\ &= k(n,p)h(n,p)[(p-1)n]^{\frac{p-1}{4}n}[2(p+1)]^{-\frac{p+1}{2}}\\ &\quad\cdot[(n+2)-p(n-2)]^{\frac{1}{4}[(n+2)-p(n-2)]}\|\varphi_n\|_1^{p-1}=\|\varphi_n\|_1^{p-1}.\end{aligned} \tag{18}$$

On the other hand, we have

$$\min_{0\neq u\in H^1(\mathbf{R}^n)} I_{p,n}(u)\leq I_{p,n}(\varphi_n)=\|\varphi_n\|_1^{p-1}. \tag{19}$$

Combining (18) and (19) yields

$$I_{p,n}(\varphi_n)=\min_{0\neq u\in H^1(\mathbf{R}^n)} I_{p,n}(u).$$

This completes the proof of Theorem 2.1. □

COROLLARY 2.5. *Let $n\geq 1$ and $1<p<\gamma(n)$. The smallest constant for which the Sobolev inequality* (3) *holds is given by* (4), *where φ_n is the ground state of equation* (5).

Remark 1. If $n=1$, the remark following Lemma 2.3 leads to

$$C_{p,1}>(p-1)^{\frac{p-1}{4(p+1)}}[2(p+1)]^{-\frac{1}{2}}(p+3)^{\frac{p+3}{4(p+1)}}e^{-\frac{p-1}{2(p+1)}}$$

and

$$C_{p,1}<(p-1)^{\frac{p-1}{4(p+1)}}(2p+2)^{-\frac{1}{2}}(p+3)^{\frac{p+3}{4(p+1)}}.$$

Remark 2. There are a lot of results in [6], [7], [14], [33], [42], [49], [39], [54], [55] on the existence of decaying positive solutions of (5). Kwong [33] proved the uniqueness of the ground state φ_n of equation (5) for $1<p<\gamma(n)$, which is positive, spherically symmetric, decreasing with respect to r, and exponentially decaying together with its derivatives up to order two.

Using Theorem 2.1, we can show the following theorem.

THEOREM 2.6.

$$\min\{E(u,0)|0\neq u\in H^1(\mathbf{R}^n), R(u)=0\}=d_n>0,$$

where $d_n=E(\varphi_n,0)$, where φ_n is the ground state of (5), $R(u)=\|u\|_1^2-|u|_{p+1}^{p+1}$, *and the energy* $E(u,v)=\frac{1}{2}\|u\|_1^2+\frac{1}{2}|v|_2^2-\frac{1}{p+1}|u|_{p+1}^{+1}$.

Proof. By (10) we have $R(\varphi_n)=0$. So

$$d_n=E(\varphi_n,0)\geq\min\{E(u,0)\mid 0\neq u\in H^1(\mathbf{R}^n), R(u)=0\}=e_n$$

and

$$d_n=E(\varphi_n,0)=\frac{1}{2}\|\varphi_n\|_1^2-\frac{1}{p+1}|\varphi|_{p+1}^{p+1}=\frac{p-1}{2(p+1)}\|\varphi_n\|_1^2>0. \tag{20}$$

On the other hand,

$$e_n = \min\left\{\frac{p-1}{2(p+1)}\|u\|_1^2 \mid 0 \neq u \in H^1(\mathbf{R}^n), R(u) = 0\right\}.$$

Let $w = \|u\|_1^{-\frac{2}{p-1}} u$. Then

$$\begin{aligned} e_n &= \frac{p-1}{2(p+1)} \min\left\{\|w\|_1^{\frac{2(p+1)}{p-1}} \mid 0 \neq w \in H^1(\mathbf{R}^n), |w|_{p+1} = 1\right\} \\ &= \frac{p-1}{2(p+1)} \min\left\{\left(\frac{\|w\|_1^{p+1}}{|w|_{p+1}^{p+1}}\right)^{\frac{2}{p-1}} \mid 0 \neq w \in H^1(\mathbf{R}^n)\right\}. \end{aligned} \tag{21}$$

By Corollary 2.5, we obtain for all $0 \neq w \in H^1(\mathbf{R}^n)$

$$\frac{\|w\|_1^{p+1}}{|w|_{p+1}^{p+1}} \geq C_{p,n}^{-(p+1)} = \alpha_{p,n} = \|\varphi_n\|_1^{p-1}.$$

Combining this estimate with (20) and (21) yields

$$e_n \geq \frac{p-1}{2(p+1)}\|\varphi_n\|_1^2 = d_n. \tag{22}$$

The proof of Theorem 2.6 is completed by (20) and (22). □

3. Invariant sets for the BQ. For $n = 1$ and $d = d_1$, recall the definition (8) of K_1 and K_2. In this section we assume $f(s) = |s|^{p-1}s$ for some $1 < p < \infty$.

LEMMA 3.1 (invariant sets). *Suppose $f(s) = |s|^{p-1}s$ with $p > 1$. Let initial data satisfy $u_0 \in K_1, v_0 \in L^2(\mathbf{R})$, and $E(u_0, v_0) < d$. Let $\vec{u}(t) = (u(t), v(t))$ be the solution of the BQ with $\vec{u}(0) = (u_0, v_0)$ such that $\vec{u} \in C([0,T), H^1 \times L^2)$ for some $T > 0$. Then $u(t) \in K_1$ for $0 \leq t < T$. On the other hand, if $u_0 \in K_2, v_0 \in L^2(\mathbf{R})$, and $E(u_0, v_0) < d$, then $u(t) \in K_2$ and $R(u(t)) < -2(d - E(u_0, v_0))$ for $0 \leq t < T$.*

Proof. We only consider the invariance of K_2 since for K_1 the proof is similar. Let $u_0 \in K_2, v_0 \in L^2$, and $E(u_0, v_0) < d$. Since $E(u(t), v(t)) = E(u_0, v_0)$ for $0 \leq t < T$, we have

$$E(u(t), 0) \leq E(u(t), v(t)) = E(u_0, v_0) < d.$$

Suppose $u(t_0) \notin K_2$ for some t_0 in $(0, T)$. That is, $R(u(t_0)) = \|u(t_0)\|_1^2 - |u(t_0)|_{p+1}^{p+1} \geq 0$. By $R(u(0)) = R(u_0) < 0$ and the continuity of $R(u(t))$ with respect to t, there exists $t_1 \in (0, t_0]$ such that $R(u(t_1)) = 0$. Theorem 2.6 yields the contradiction

$$d > E(u(t_1), 0) \geq \min\{E(u, 0); 0 \neq u \in H^1(\mathbf{R}), R(u) = 0\} = d.$$

Thus $u(t) \in K_2$ for $0 \leq t < T$.

To prove the final inequality, we define the function $W(\rho) = R(\rho u)$ for $\rho > 0$, where $u = u(t)$ is given above and t is fixed. Observe that $W(1) = R(u) < 0$ and

$$W(\rho) = \rho^2\|u\|_1^2 - \rho^{p+1}|u|_{p+1}^{p+1} > 0$$

for ρ sufficiently small. Hence there exists some $\rho_0 \in (0,1)$ such that $W(\rho_0) = R(\rho_0 u) = 0$. That is, $\rho_0^2\|u\|_1^2 = \rho_0^{p+1}|u|_{p+1}^{p+1}$. By Theorem 2.6, we obtain

$$\begin{aligned} d &= \min\{E(u,0); 0 \neq u \in H^1(\mathbf{R}), R(u) = 0\} \leq E(\rho_0 u, 0) \\ &= \frac{1}{2}\,\rho_0^2\|u\|_1^2 - \frac{1}{p+1}\,\rho_0^{p+1}|u|_{p+1}^{p+1} \\ &= \left(\frac{1}{2} - \frac{1}{p+1}\right)\rho_0^{p+1}|u|_{p+1}^{p+1} < \left(\frac{1}{2} + \frac{1}{p+1}\right)|u|_{p+1}^{p+1} \\ &= \frac{1}{2}\,\|u\|_1^2 - \frac{1}{p+1}|u|_{p+1}^{p+1} - \frac{1}{2}\,\left(\|u\|_1^2 - |u|_{p+1}^{p+1}\right) \\ &= E(u,0) - \frac{1}{2}\,R(u) \leq E(u,v) - \frac{1}{2}\,R(u) = E(u_0,v_0) - \frac{1}{2}\,R(u). \end{aligned}$$

That is,

$$R(u(t)) < 2(E(u_0,v_0) - d) \quad \text{for } 0 \leq t < T.$$

This completes the proof of Lemma 3.1. □

COROLLARY 3.2. *Let $f(s) = |s|^{p-1}s$ with $p > 1$. If $u_0 \in K_2, v_0 \in L^2$, and $E(u_0,v_0) < d$, then $|u(t)|_{p+1} + \frac{p+1}{p-1}|v(t)|_2^2 > |\varphi|_{p+1}$ for all $t \in [0,T)$. If $u_0 \in K_1$, then $\|u(t)\|_1 < \|\varphi\|_1 |u(t)|_{p+1} < |\varphi|_{p+1}$, and $E(u(t),v(t)) < 0$ for all $t \in [0,T)$.*

Proof. Let $u_0 \in K_2, v_0 \in L^2$, and $E(u_0,v_0) < d$. By Lemma 3.1 and

$$d = E(\varphi,0) = \frac{p-1}{2(p+1)}|\varphi|_{p+1}^{p+1},$$

we have

$$\begin{aligned} \|u(t)\|_1^2 - |u(t)|_{p+1}^{p+1} &= R(u(t)) < 2(E(u(t),v(t)) - d) \\ &= \|u(t)\|_1^2 - \frac{2}{p+1}|u(t)|_{p+1}^{p+1} + |v(t)|_2^2 - \frac{p-1}{p+1}|\varphi|_{p+1}^{p+1}. \end{aligned}$$

This implies $|u(t)|_{p+1} + \frac{p+1}{p-1}|v(t)|_2^2 > |\varphi|_{p+1}$. On the other hand, if $u_0 \in K_1, v_0 \in L^2$, and $E(u_0,v_0) < d$, by Lemma 3.1 we have $R(u(t)) > 0$ and $E(u(t),v(t)) < d$ for all $t \in [0,T)$. This implies

$$\frac{p-1}{2(p+1)}\|\varphi\|_1^2 = d > E(u,v) > \frac{1}{2}\|u(t)\|_1^2 - \frac{1}{p+1}\|u(t)\|_1^2 = \frac{p-1}{2(p+1)}\|u(t)\|_1^2.$$

That is, $\|u(t)\|_1 < \|\varphi\|_1$, so that

$$|u(t)|_{p+1} < \|u(t)\|_1^{\frac{2}{p+1}} < \|\varphi\|_1^{\frac{2}{p+1}} = |\varphi|_{p+1}.$$

This completes the proof of Corollary 3.2. □

Remark. If $E(u_0,v_0) \leq 0$, we have $u_0 \in K_2$.

The following result was obtained by Liu [37].

THEOREM 3.3 (local existence). *Let $\vec{u}_0 = (u_0,v_0) \in H^1(\mathbf{R}) \times L^2(\mathbf{R})$; then there exist $T > 0$ and a unique weak solution $\vec{u} = (u,v)$ of the BQ in $C([0,T); H^1 \times L^2)$ with $\vec{u}(0) = \vec{u}_0$ such that $E(\vec{u}) = E(\vec{u}_0), V(\vec{u}) = V(\vec{u}_0), I_1(\vec{u}) = I_1(\vec{u}_0)$, and $I_2(\vec{u}) = I_2(\vec{u}_0)$.*

Furthermore, the interval of existence $[0,T)$ *can be extended to a maximal interval* $[0,T_{\max})$ *such that either*

(i) $T_{\max}=+\infty$, *or*

(ii) $T_{\max}<+\infty$, $\lim_{t\to T_{\max}}\|\vec{u}(t)\|_{H^1\times L^2}=+\infty$.

Remark 1. For a general function $f\in C^{s+1}(\mathbf{R})$ with $f(0)=0$ and initial data $\vec{u}_0\in H^{s+2}\times H^{s+1}$ for sufficiently large s (for example $s>\frac{5}{2}$), we have a unique, local classical solution to the BQ, which lies in a Sobolev space of high order. This can be proved using the results of Bona and Sachs [8] or Kato [30].

Remark 2. In Theorem 3.3, if we further assume $\xi^{-1}\hat{u}_0\in L^2$, then

$$\xi^{-1}\hat{u}\in C^1([0,T_{\max});L^2),$$

where $\hat{u}$ is the Fourier transform of u. In fact, by the BQ, we have $\xi^{-1}\hat{u}(t,\xi)=\xi^{-1}\hat{u}_0(\xi)+\int_0^t\hat{v}(\tau,\xi)\,d\tau$ and the statement follows.

THEOREM 3.4 (global existence in K_1). *Let* $f(s)=|s|^{p-1}s$ *with* $p>1$. *If* $u_0\in K_1, v_0\in L^2$, *and* $E(u_0,v_0)<d$, *then there exists a unique global weak solution* $\vec{u}=(u,v)$ *of the BQ* (2) *in* $C([0,\infty);H^1\times L^2)$ *with* $\vec{u}(0)=\vec{u}_0$ *and* $E(\vec{u}),V(\vec{u}),I_1(\vec{u})$, *and* $I_2(\vec{u})$ *are invariant for all* $t>0$.

Proof. It suffices to prove the a priori estimate $\|u(t)\|_1+|v(t)|_2\le C(T_{\max})$ for all $0\le t<T_{\max}$. By Corollary 3.2, we obtain $u(t)\in K_1$ and $\|u(t)\|_1<\|\varphi\|_1$ for $t\in[0,T_{\max}$. By the conserved energy $E(\vec{u}(t))=E(\vec{u}_0)$, we obtain

$$\begin{aligned}\frac{1}{2}|v(t)|_2^2&\le\frac{1}{p+1}|u(t)|_{p+1}^{p+1}+E(u_0,v_0)<\frac{1}{p+1}C_{p,1}^{p+1}\|u(t)\|_1^{p+1}+d\\&<\frac{1}{p+1}\|\varphi\|_1^2+\frac{p-1}{2(p+1)}\|\varphi\|_1^2=\frac{1}{2}\|\varphi\|_1^2.\end{aligned}$$

This completes the proof of Theorem 3.4. □

4. Finite blow-up time. The following blow-up theorem is a variation of the theorem of Levine [34].

THEOREM 4.1 (blow-up). *Suppose that*

$$sf(s)\ge(2+\varepsilon)F(s)\quad\text{for all } s\in\mathbf{R}\tag{23}$$

for some $\varepsilon>0$, *where* $F'=f$ *with* $F(0)=0$. *Let* $u_0\in H^1(\mathbf{R}), v_0\in L^2(\mathbf{R})$, *and* $\xi^{-1}\hat{u}_0\in L^2(\mathbf{R})$. *Assume one of the following conditions.*

(i) $E(u_0,v_0)<0$, *or*

(ii) $E(\vec{u}_0)\ge 0$ *and* $(E(\vec{u}_0))^{\frac{1}{2}}<\dfrac{1}{\sqrt{2}}\dfrac{\Re\langle\xi^{-1}\hat{u}_0,\hat{v}_0\rangle}{|\xi^{-1}\hat{u}_0|_2}$,

where $\Re\int_{-\infty}^{\infty}\xi^{-1}\hat{u}_0\bar{\hat{v}}_0\,d\xi=\Re\langle\hat{v}_0,\xi^{-1}\hat{u}_0\rangle$. *If* $\vec{u}=(u,v)$ *is the solution of the BQ with* $\vec{u}(0)=(u_0,v_0)$ *such that*

$$\vec{u}\in C([0,T_{\max});H^1\times L^2),$$

where $T_{\max}$ *is the maximal existence time of* $\vec{u}$, *then* $T_{\max}<+\infty$ *and*

$$\lim_{t\to T_{\max}^-}(\|u(t)\|_1+|v(t)|_2)=+\infty.$$

Proof. To establish the proofs of (i) and (ii) with $E(\vec{u}_0) = 0$, consider the function of t,

$$I(t) = |\xi^{-1}\hat{u}|_2^2 + \beta(t+\tau)^2, \tag{24}$$

where β and τ are nonnegative constants, which we shall determine later, and $\hat{u}$ is the Fourier transform of u. If $T_{\max} = +\infty$, we will obtain a contradiction. Notice that

$$I'(t) = 2\Re\langle \xi^{-1}\hat{u}, \xi^{-1}\hat{u}_t\rangle + 2\beta(t+\tau) \tag{25}$$

and

$$\begin{aligned} I''(t) &= 2|\hat{v}(t)|_2^2 - 2\Re\langle \xi^{-1}\hat{u}, \xi(\hat{u} + \xi^2\hat{u} - \hat{f}(u))\rangle + 2\beta \\ &= 2|v(t)|_2^2 - 2\|u(t)\|_1^2 + 2\langle u, f(u)\rangle + 2\beta. \end{aligned} \tag{26}$$

On the other hand, using conserved energy $E(\vec{u}) = E(\vec{u}_0)$ and the condition (23), we obtain

$$\begin{aligned} I''(t) &= 4(1+\alpha)|v(t)|_2^2 + 4\alpha\|u(t)\|_1^2 - 4(1+2\alpha)E(\vec{u}_0) \\ &\quad + 2\beta + 2\int_{-\infty}^{\infty} (uf(u) - (2+4\alpha)F(u))\,dx \\ &\geq 4(1+\alpha)|v(t)|_2^2 + 4\alpha\|u(t)\|_1^2 - 4(1+2\alpha)E(u_0, v_0) + 2\beta, \end{aligned} \tag{27}$$

where $\alpha = \frac{\epsilon}{4} > 0$. It follows that

$$II'' - (1+\alpha)(I')^2 \geq 4(1+\alpha)S(t) + 2I\left(2\alpha\|u\|_1^2 - 2(1+2\alpha)\left(E(\vec{u}_0) - \frac{1}{2}\,\beta\right)\right), \tag{28}$$

where

$$S(t) = (|\xi^{-1}\hat{u}|_2^2 + \beta(t+\tau)^2)(|v(t)|_2^2 + \beta) - (\Re\langle \xi^{-1}\hat{u}, \hat{v}\rangle + \beta(t+\tau))^2 \geq 0.$$

Thus

$$II'' - (1+\alpha)(I')^2 > -2(1+2\alpha)(2E(u_0, v_0) + \beta)I. \tag{29}$$

Now suppose (i) holds. Then taking $\beta = -2E(\vec{u}_0)$, we find the inequality

$$I(t)I''(t) - (1+\alpha)(I'(t))^2 > 0. \tag{30}$$

Choose τ so large that $I'(0) = 2\Re\langle \xi^{-1}\hat{u}_0, v_0\rangle - 4E(\vec{u}_0)\tau > 0$. Hence $J(t) = [I(t)]^{-\alpha}$ satisfies $J''(t) = -\alpha I^{-\alpha-2}(II'' - (1+\alpha)(I')^2 < 0$ for all $t \geq 0$, as well as $J(0) > 0$ and $J'(0) < 0$. Hence $J(t) \leq J(0) + tJ'(0)$, so $J(t_0) = 0$, where $0 < t_0 \leq \frac{I(0)}{\alpha I'(0)} < +\infty$. Thus we see that

$$\lim_{t \to t_0^-} |\xi^{-1}\hat{u}(t)|_2 = \lim_{t \to t_0^-} [J(t)]^{-\frac{1}{2\alpha}} = +\infty. \tag{31}$$

If (ii) holds with $E(u_0, v_0) = 0$, let $\beta = 0$, so that $J''(t) = (I^{-\alpha}(t))'' < 0, J'(0) = -\alpha I^{-\alpha-1}(0)I'(0) < 0$, and $J(0) > 0$. Therefore, we obtain (31) in this case as well.

Now if (ii) holds with $E(u_0, v_0) > 0$, let $\beta = 0$. Then

$$II'' - (1+\alpha)(I')^2 \geq -4(1+2\alpha)E(\vec{u}_0)I \tag{32}$$

and

$$J'(0) = -\alpha I^{-\alpha-1}(0)I'(0) < 0.$$

By the continuity of $J'(t)$, there exists $\eta > 0$ such that $J'(t) < 0$ for $t \in [0, \eta)$. Let

$$t^* = \sup\{t \mid J'(\tau) < 0 \quad \text{for } \tau \in [0, t)\}.$$

We have $0 < t^* \leq T_{\max}$, where $T_{\max}$ is the maximal existence time. Multiplying (32) by $-\alpha I^{-\alpha-2}(t)J'(t)$ for $t \in [0, t^*)$, we obtain

$$[(J'(t))^2]' \geq 8\alpha^2 E(\vec{u}_0)(I^{-2\alpha-1}(t))' \quad \text{for } t \in [0, t^*). \tag{33}$$

Integrating (33) in $[0, t)$, we obtain

$$(J'(t))^2 \geq (J'(t))^2 - 8\alpha^2 E(\vec{u}_0)I^{-2\alpha-1}(t) \geq (J'(0))^2 - 8\alpha^2 E(\vec{u}_0)I^{-2\alpha-1}(0). \tag{34}$$

Assumption (ii) implies that

$$J'(0) + \sqrt{8}\alpha(E(\vec{u}_0))^{\frac{1}{2}} I^{-\alpha-\frac{1}{2}}(0) < 0.$$

So

$$(J'(0))^2 - 8\alpha^2 E(\vec{u}_0)I^{-2\alpha-1}(0) > 0.$$

Hence by the continuity of $J'(t)$ and $J'(0)$, we obtain $J'(t) < 0$ for $t \in [0, t^*)$. Moreover,

$$J'(t) \leq -((J'(0))^2 - 8\alpha^2 E(\vec{u}_0)I^{-2\alpha-1}(0))^{\frac{1}{2}} \equiv -\alpha < 0 \quad \text{for } 0 \leq t < t^*.$$

It follows that $t^* = T_{\max}$ and $J'(t) \leq -a$ for all $t \geq 0$. Therefore, $J(t) \leq J(0) - at$ for all $t \geq 0$. So $J(t_0) = 0$, where $0 < t_0 \leq \frac{J(0)}{a}$. Hence in all cases we obtain

$$\lim_{t \to t_0} I(t) = \lim_{t \to t_0} |\xi^{-1}\hat{u}(t)|_2^2 = +\infty.$$

Using the BQ, we have $\hat{u}_t = i\xi\hat{v}$ and

$$|\xi^{-1}\hat{u}(t)|_2 \leq |\xi^{-1}\hat{u}_0|_2 + \int_0^t |\hat{v}(\tau)|_2 \, d\tau.$$

Letting $t \to t_0^-$, we obtain

$$\int_0^{t_0} |\hat{v}(\tau)|_2 \, d\tau = +\infty.$$

This implies that there exists a sequence $\{t_n\}, 0 < t_n < t_0$, such that

$$\lim_{t_n \to t_0^-} |v(t_n)|_2 = \lim_{t_n \to t_0^-} |\hat{v}(t_n)|_2 = +\infty.$$

This contradicts $T_{\max} = +\infty$. By the local existence theorem (Theorem 3.3), we deduce $T_{\max} \leq t_0 < +\infty$ and $\lim_{t \to T_{\max}^-}(\|u(t)\|_1 + |v(t)|_2) = +\infty$. This completes the proof of Theorem 4.1. □

Remark 1. The special case $f(s) = |s|^{p-1}s$ with $p > 1$ satisfies assumption (23). In that case, under the conditions of Theorem 4.1, we have $T_{\max} < \infty$ and

$$\lim_{t \to T_{\max}^-} \|u(t)\|_1 = \lim_{t \to T_{\max}^-} |u(t)|_{p+1} = +\infty.$$

In fact, since $T_{\max} \le t_0 < +\infty$ and $\lim_{t \to T_{\max}^-} (\|u(t)\|_1 + |v(t)|_2) = +\infty$, we obtain either

$$\lim_{t \to T_{\max}^-} \|u(t)\|_1 = +\infty \quad \text{or} \quad \lim_{t \to T_{\max}^-} |v(t)|_2 = +\infty.$$

In the latter case, we have

$$\frac{1}{2}|v(t)|_2^2 \le E(\vec{u}_0) + \frac{1}{p+1}|u(t)|_{p+1}^{p+1} \le |E(\vec{u}_0)| + C_{p,1}^{p+1}\|u(t)\|_1^{p+1}.$$

Hence

$$\lim_{t \to T_{\max}^-} |u(t)|_{p+1} = \lim_{t \to T_{\max}^-} \|u(t)\|_1 = +\infty.$$

In either case, we obtain

$$\lim_{t \to T_{\max}^-} \|u(t)\|_1 = +\infty.$$

On the other hand, by $E(\vec{u}(t)) = E(\vec{u}_0)$, we have

$$\frac{1}{2}\|u(t)\|_1^2 \le E(\vec{u}_0) + \frac{1}{p+1}|u(t)|_{p+1}^{p+1}.$$

Hence,

$$\lim_{t \to T_{\max}^-} |u(t)|_{p+1} = +\infty.$$

Remark 2. If $f(s) = |s|^{p-1}s$ with $p > 1$, assumption (ii) in Theorem 4.1 implies that $R(u(t)) < 0$, for all $t \in [0, T_{\max})$. In fact, we have

$$E(\vec{u}_0) < \frac{1}{2}\,\frac{(\Re\langle \xi^{-1}\hat{u}_0, \hat{v}_0\rangle)^2}{|\xi^{-1}\hat{u}_0|_2^2} \le \frac{1}{2}\,|v_0|_2^2.$$

This implies

$$\frac{1}{2}\,R(u(t)) < \frac{1}{2}\,\|u(t)\|_1^2 - \frac{1}{p+1}|u(t)|_{p+1}^{p+1} < 0 \quad \text{for } t \in [0, T_{\max}).$$

Now we use our "best constant" results in §2 to obtain a more general blow-up result, in which the energy could be larger.

THEOREM 4.2 (improved blow-up). *Let $f(s) = |s|^{p-1}s$ with $p > 1$. Assume*

(i) $v_0 \in L^2(\mathbf{R}), u_0 \in H^1(\mathbf{R})$, *and* $\xi^{-1}\hat{u}_0 \in L^2(\mathbf{R})$;

(ii) $u_0 \in K_2$ *and* $E(u_0, v_0) < d$.

Let $\vec{u} = (u, v)$ be the solution of the BQ with $\vec{u}(0) = (u_0, v_0)$ such that $\vec{u} \in C([0, T_{\max}); H^1 \times L^2)$, where $T_{\max}$ is the maximum existence time. Then $T_{\max} < +\infty$ and

$$\lim_{t \to T_{\max}^-} \|u(t)\|_1 = \lim_{t \to T_{\max}^-} |u(t)|_{p+1} = +\infty.$$

Proof. If $T_{\max} = +\infty$, we will get a contradiction. Recall identities (26) and (27) for $\beta = 0$ and $\alpha = \frac{p-1}{4}$, which are $I(t) = |\xi^{-1}\hat{u}(t)|_2^2$,

$$I''(t) = 2|v(t)|_2^2 - 2R(v(t)), \tag{35}$$

and

$$\begin{aligned} I''(t) &= (p+3)|v(t)|_2^2 + (p-1)\|u(t)\|_1^2 - 2(p+1)E(u_0, v_0) \\ &\equiv (p+3)|v(t)|_2^2 + M(t). \end{aligned} \tag{36}$$

By Lemma 3.1, assumption (ii) means $R(u(t)) < 0$ for all $t \geq 0$. Thus

$$\|u(t)\|_1^2 < |u(t)|_{p+1}^{p+1} \leq C_{p,1}^{p+1}\|u(t)\|_1^{p+1}.$$

By Corollary 2.5, this implies

$$\|u(t)\|_1^2 > C_{p,1}^{-\frac{2(p+1)}{p-1}} = \|\varphi\|_1^2$$

and

$$M(t) > (p-1)\left[\|u(t)\|_1^2 - \frac{2(p+1)}{p-1}\, d\right] \geq (p-1)[\|\varphi\|_1^2 - \|\varphi\|_1^2] = 0.$$

Hence $I''(t) > (p+3)|v(t)|_2^2$. Using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} I(t)I''(t) - \frac{p+3}{4}(I'(t))^2 &> (p+3)(|\xi^{-1}\hat{u}|_2^2|v|_2^2 - (\Re\langle\xi^{-1}\hat{u}, \xi^{-1}\hat{u}_t\rangle)^2) \\ &= (p+3)(|\xi^{-1}\hat{u}|_2^2|v|_2^2 - (\Re\langle\xi^{-1}\hat{u}, \hat{v}\rangle)^2) \geq 0. \end{aligned} \tag{37}$$

We claim that $I'(t_2) > 0$ for some $t_2 > 0$. Assuming the claim, define $J(t) = [I(t)]^{-\alpha}$, where $\alpha = \frac{p-1}{4}$. Inequality (37) means $J''(t) > 0$ for all $t \geq 0$.

Suppose that $I'(t_2) > 0$. Then $J'(t_2) < 0$, for some $t_2 > 0$, that is, $I'(t_2) > 0$, so there exists $t_3 \in \left(0, -\frac{J(t_2)}{J'(t_2)}\right]$ such that $J(t_3) = 0$. Hence

$$\lim_{t \to t_3^-} |\xi^{-1}\hat{u}(t)|_2^2 = \lim_{t \to t_3^-} I(t) = +\infty.$$

Hence there exists a sequence $\{t_n\} \to t_3^-$ such that $\lim |v(t_n)|_2 = +\infty$. This can be done by following the proof of Theorem 4.1. Thus we get a contradiction with $T_{\max} = +\infty$. Therefore, $T_{\max} < \infty$ and

$$\lim_{t \to T_{\max}^-} (\|u(t)\|_1 + |v(t)|_2) = +\infty.$$

Because $E(\vec{u}(t)) = E(\vec{u}_0)$, it follows that

$$\lim_{t \to T_{\max}^-} |u(t)|_{p+1} = \lim_{t \to T_{\max}^-} \|u(t)\|_1 = +\infty.$$

Now we prove the claim that $I'(t_2) > 0$ for some $t_2 > 0$. If not, then $I'(t) \leq 0$ for all $t \geq 0$. By (35) and $R(u(t)) < 0$,

$$I''(t) = 2|v(t)|_2^2 - 2R(u(t)) > 2|v(t)|_2^2 \geq 0; \tag{38}$$

the limit

$$\lim_{t\to+\infty} I'(t) = I'(0) + \int_0^{+\infty} I''(t)\,dt$$

exists. Hence there exists a sequence $\{t_n\}$ such that

$$\lim_{t_n\to+\infty} I''(t_n) = 0.$$

By (38), we obtain

$$\lim_{t_n\to+\infty} |v(t_n)|_2 = \lim_{t_n\to+\infty} R(u(t_n)) = 0$$

and

$$\lim_{t_n\to+\infty} E(u(t_n), 0) = \lim_{t_n\to\infty} E(u(t_n), v(t_n)) = E(u_0, v_0) < d.$$

On the other hand, from Lemma 3.1 we have

$$2(E(u_0, v_0) - d) > R(u(t_n)) \to 0 \quad \text{as } t_n \to +\infty.$$

This implies that $E(u_0, v_0) \geq d$. This contradicts assumption (ii). □

As a result of Theorem 4.2, the following instability theorem follows. It asserts that there are solutions with initial data arbitrarily near the ground state that blow up in a finite time.

THEOREM 4.3 (instability theorem). *Let $f(s) = |s|^{p-1}s$ with $p > 1$. Let $\varphi \in H^1(\mathbf{R})$ be the ground state of (5). For any $\delta > 0$, there is initial data $u_0 \in H^1(\mathbf{R})$ with $\|u_0 - \varphi\|_1 < \delta$, such that the solution $\vec{u} = (u, v)$ of the BQ with $\vec{u}(0) = (u_0, 0)$ satisfies*

$$\lim_{t\to T^-} \|u(t)\|_1 = +\infty$$

for some $0 < T < +\infty$.

To prove Theorem 4.3, we need the following lemma.

LEMMA 4.4. *The set $A = \{w \in H^1(\mathbf{R}) \mid \xi^{-1}\hat{w}(\xi) \in L^2(\mathbf{R})\}$ is dense in $H^1(\mathbf{R})$.*

Proof. For any $u \in H^1(\mathbf{R})$ and $\varepsilon > 0$, define w_δ such that

$$\hat{w}_\delta(\xi) = \begin{cases} \hat{u}(\xi), & |\xi| > \delta, \\ 0, & |\xi| \leq \delta. \end{cases}$$

Then

$$|\xi^{-1}\hat{w}_\delta|_2^2 = \int_{|\xi|>\delta} \xi^{-2}|\hat{u}(\xi)|^2\,d\xi < \delta^{-2}|u|_2^2 < +\infty.$$

and

$$\|w_\delta\|_1 = |(1+\xi^2)^{\frac{1}{2}}\hat{w}_\delta|_2 \leq |(1+\xi^2)^{\frac{1}{2}}\hat{u}(\xi)|_2 = \|u\|_1 < +\infty.$$

This implies $w_\delta \in A$. On the other hand, we have

$$\begin{aligned} \|w_\delta - u\|_1^2 &= |(1+\xi^2)^{\frac{1}{2}}(\hat{w}_\delta - \hat{u})|_2^2 \\ &= \int_{|\xi|\leq\delta} (1+\xi^2)|\hat{u}|^2\,d\xi \leq \|u\|_1^2 < +\infty. \end{aligned}$$

Hence we can choose δ to be sufficiently small so that

$$\int_{|\xi|\leq\delta} (1+\xi^2)|\hat{u}|^2\,d\xi < \varepsilon.$$

This completes the proof of Lemma 4.4. □

Proof of Theorem 4.3. For any $\delta > 0$, let $0 < \varepsilon_0 < \min\{\frac{\delta}{2}, \|\varphi\|_1\}$ and $\varepsilon_1 < \frac{\delta}{4\|\varphi\|_1}$. By Lemma 4.4, there exists $w_0 \in A$ such that $\|w_0 - \varphi\|_1 < \varepsilon_0$. Let $u_0 = (1+\varepsilon_1)w_0$ so that $u_0 \in A$. Then

$$\|u_0 - \varphi\|_1 < \varepsilon_0 + 2\|\varphi\|_1\varepsilon_1 < \delta.$$

By Theorem 4.2, it suffices to show that $E(u_0, 0) < d$ and $R(u_0) < 0$. Using the "best constant" result (Corollary 2.5), we have

$$\begin{aligned} |w_0|_{p+1} &\geq |\varphi|_{p+1} - |w_0 - \varphi|_{p+1} \geq |\varphi|_{p+1} - \|\varphi\|_1^{-\frac{p-1}{p+1}}\|w_0 - \varphi\|_1 \\ &> |\varphi|_{p+1} - \|\varphi\|_1^{-\frac{p-1}{p+1}}\varepsilon_0. \end{aligned}$$

But

$$|\varphi|_{p+1} - \|\varphi\|_1^{-\frac{p-1}{p+1}}\varepsilon_0 > |\varphi|_{p+1} - \|\varphi\|_1^{-\frac{p-1}{p+1}}\|\varphi\|_1 = |\varphi|_{p+1} - \|\varphi\|_1^{\frac{2}{p+1}} = 0.$$

This implies

$$|w_0|_{p+1}^{p+1} > \left(|\varphi|_{p+1} - \|\varphi\|_1^{-\frac{p-1}{p+1}}\varepsilon_0\right)^{p+1} = |\varphi|_{p+1}^{p+1} + O(\varepsilon_0).$$

Next,

$$\|w_0\|_1^2 \leq (\|w_0 - \varphi\|_1 + \|\varphi\|_1)^2 < (\varepsilon_0 + \|\varphi\|_1)^2 = \|\varphi\|_1^2 + O(\varepsilon_0).$$

Therefore, we obtain

$$\begin{aligned} E(u_0, 0) &= \frac{1}{2}(1+\varepsilon_1)^2\|w_0\|_1^2 - \frac{1}{p+1}(1+\varepsilon_1)^{p+1}|w_0|_{p+1}^{p+1} \\ &< \frac{1}{2}(1+\varepsilon_1)^2(\|\varphi\|_1^2 + O(\varepsilon_0)) - \frac{1}{p+1}(1+\varepsilon_1)^{p+1}(|\varphi|_{p+1}^{p+1} + O(\varepsilon_0)) \\ &= h(\varepsilon_1)\|\varphi\|_1^2 + O(\varepsilon_0) \end{aligned}$$

since $\|\varphi\|_1^2 = |\varphi|_{p+1}^{p+1}$, where $h(\varepsilon_1) = \frac{1}{2}(1+\varepsilon_1)^2 - \frac{1}{p+1}(1+\varepsilon_1)^{p+1}$. Since

$$h'(a) = (1+a)[1-(1+a)^p] < 0 \text{ for } a > 0,$$

we have

$$h(\varepsilon_1)\|\varphi\|_1^2 < h(0)\|\varphi\|_1^2 = \left(\frac{1}{2} - \frac{1}{p+1}\right)\|\varphi\|_1^2 = d.$$

Choose ε_0 sufficiently small so that $O(\varepsilon_0) < d - h(\varepsilon_1)\|\varphi\|_1^2$. Then

$$E(u_0, 0) < h(\varepsilon_1)\|\varphi\|_1^2 + O(\varepsilon_0) < d.$$

On the other hand,

$$\begin{aligned} R(u_0) &= (1+\varepsilon_1)^2\|w_0\|_1^2 - (1+\varepsilon_1)^{p+1}|w_0|_{p+1}^{p+1} \\ &< (1+\varepsilon_1)^2(\|\varphi\|_1^2 + O(\varepsilon_0)) - (1+\varepsilon_1)^{p+1}(|\varphi|_{p+1}^{p+1} + O(\varepsilon_0)) \\ &= ((1+\varepsilon_1)^2 - (1+\varepsilon_1)^{p+1})\|\varphi\|_1^2 + O(\varepsilon_0). \end{aligned}$$

Finally, choose $\varepsilon_0 > 0$ sufficiently small so that

$$O(\varepsilon_0) < ((1+\varepsilon_1)^{p+1} - (1+\varepsilon_1)^2)\|\varphi\|_1^2.$$

Then $R(u_0) < 0$. This completes the proof of Theorem 4.3. □

5. Global solution for the solitary wave case. In this section, we will prove a more general global existence theorem for the BQ. It turns out that such a solution is bounded by the solitary wave solution φ_c, which is defined in (39). The proof is similar to that in §3.

Define

$$K_1^c = \{u \in H^1(\mathbf{R}) | L_c(u, -cu) < d(c), R_c(u) > 0\},$$
$$K_2^c = \{u \in H^1(\mathbf{R}) | L_c(u, -cu) < d(c), R_c(u) < 0\},$$

where $L_c(u,v) = E(u,v) + cV(u,v), d(c) = L_c(\varphi_c, -c\varphi_c), R_c(u) = (1-c^2)|u|_2^2 + |u_x|_2^2 - |u|_{p+1}^{p+1}$, and φ_c is the ground state of

$$-u_{xx} + (1-c^2)u - |u|^{p-1}u = 0 \tag{39}$$

for $|c| < 1$. Also, $\varphi_c(\xi) = \varphi_c(x-ct) = \alpha_1 \operatorname{sech}^{\frac{2}{p-1}}(\alpha_2 \xi)$ is the solitary wave solution of the BQ, where $\alpha_1 = \left[\frac{1}{2}(p+1)(1-c^2)\right]^{\frac{1}{p-1}}$ and $\alpha_2 = \frac{1}{2}(1-c^2)^{\frac{1}{2}}(p-1)$.

THEOREM 5.1 (global existence theorem in K_1^c). *Let $f(s) = |s|^{p-1}s$ with $p > 1$. Let $|c| < 1, u_0 \in K_1^c, v_0 \in L^2$, and $L_c(u_0, v_0) < d(c)$, where $L_c(\vec{u}) = E(\vec{u}) + cV(\vec{u})$ and $d(c) = L_c(\vec{\varphi}_c) = L_c(\varphi_c, -c\varphi_c)$. Then there exists a unique global weak solution $\vec{u} = (u, v)$ of the BQ in $C([0,\infty); H^1 \times L^2)$ with $\vec{u}(0) = \vec{u}_0$ such that $E(\vec{u}), V(\vec{u}), I_1(\vec{u})$, and $I_2(\vec{u})$ are invariant for all $t \geq 0$.*

To prove Theorem 5.1, we need some lemmas.

LEMMA 5.2. *Let $|c| < 1$ and $d(c) = L_c(\varphi_c, -c\varphi_c)$; then*

$$M_c \equiv \min\{L_c(u, -cu) | 0 \neq u \in H^1(\mathbf{R}), R_c(u) = 0\} = d(c) > 0$$

where φ_c is the ground state of equation (39), *satisfying*

$$-\varphi_c'' + (1-c^2)\varphi_c - |\varphi_c|^{p-1}\varphi_c = 0,$$

$$R_c(u) = (1-c^2)|u|_2^2 + |u_x|_2^2 - |u|_{p+1}^{p+1},$$

and

$$L_c(u, -cu) = E(u, -cu) + cV(u, -cu) = \frac{1}{2}(1-c^2)|u|_2^2 + \frac{1}{2}|u_x|_2^2 - \frac{1}{p+1}|u|_{p+1}^{p+1}.$$

Proof. It is easy to see $R_c(\varphi_c) = 0$. Hence, $d(c) \geq M_c$. On the other hand, for any nonzero $u \in H^1$, with $R_c(u) = 0$, let $u(x) = \lambda^{\frac{2}{p-1}} w(\lambda x)$ with $\lambda = (1-c^2)^{\frac{1}{2}}$. We have

$$R_c(u) = \lambda^{\frac{p+3}{p-1}}(\|w\|_1^2 - |w|_{p+1}^{p+1}) = \lambda^{\frac{p+3}{p-1}} R(w)$$

and

$$L_c(u, -cu) = \lambda^{\frac{p+3}{p-1}}\left(\frac{1}{2}\|w\|_1^2 - \frac{1}{p+1}|w|_{p+1}^{p+1}\right) = \lambda^{\frac{p+3}{p-1}} E(w, 0).$$

So by Theorem 2.6,

$$\begin{aligned} L_c(u,-cu) &\geq \lambda^{\frac{p+3}{p-1}} \min\{E(v,0) | 0 \neq v \in H^1, R(v) = 0\} \\ &= \lambda^{\frac{p+3}{p-1}} E(\varphi, 0) = \lambda^{\frac{p+3}{p-1}} d, \end{aligned}$$

but $\varphi_c(x) = \lambda^{\frac{2}{p-1}} \varphi(\lambda x)$ and $d(c) = \lambda^{\frac{p+3}{p-1}} d > 0$, where φ is the ground state of (5). With this fact, it is observed that $d(c) \leq M_c$. Therefore, $d(c) = M_c$. □

LEMMA 5.3 (invariant sets). *Suppose $f(s) = |s|^{p-1}s$ with $p > 1$. Let $|c| < 1$ and the initial data $u_0 \in K_1^c, v_0 \in L^2(\mathbf{R})$, and $L_c(u_0, v_0) < d(c)$. Let $\vec{u} = (u, v)$ be the solution of the BQ with $\vec{u}(0) = (u_0, v_0)$ such that $\vec{u} \in C([0,T), H^1 \times L^2)$ for some $T > 0$. Then $u(t) \in K_1^c$ for $0 \leq t < T$. On the other hand, if $u_0 \in K_2^c, v_0 \in L^2(\mathbf{R})$, and $L_c(u_0, v_0) < d(c)$, then $u(t) \in K_2^c$ and $R_c(u(t)) < -2(d(c) - L_c(u_0, v_0))$ for $0 \leq t < T$.*

Proof. We only prove the invariance of K_2^c since the proof of the K_1^c case is similar. Let $u_0 \in K_2^c$ for $|c| < 1, v_0 \in L^2$, and $L_c(u_0, v_0) < d(c)$. By $L_c(u(t), v(t)) = \text{const}$,

$$\begin{aligned} L_c(u(t), -cu(t)) &= L_c(u(t), v(t)) - \left(\frac{1}{2} c^2 |u(t)|_2^2 + \frac{1}{2} |v(t)|_2^2 + cV(u(t), v(t)) \right) \\ &\leq L_c(u(t), v(t)) = L_c(u_0, v_0) < d(c) \end{aligned}$$

for $0 \leq t < T$. If $u(t) \notin K_2^c$ for some t in $(0,T)$, by the continuity of $R_c(u(t))$ with respect to t in $[0,T)$, there exists t_0 in $(0,T)$ such that $R(u(t_0)) = 0$. So it yields the contradiction

$$\begin{aligned} d(c) &> L_c(u(t_0), -cu(t_0)) \\ &\geq \min\{L_c(u, -cu) | 0 \neq u \in H^1(\mathbf{R}), R_c(u) = 0\} = d(c). \end{aligned}$$

This proves $u(t) \in K_2^c$ for all $t \in [0,T)$. For the rest of the theorem, we define the function $W_c(\rho) = R_c(\rho u)$ with $\rho > 0$, where $|c| < 1$ and $u \in K_2^c$ is the solution of the BQ with $\vec{u}(0) = \vec{u}_0$. Then $W_c(1) = R_c(u) < 0$ and

$$W_c(\rho) = \rho^2((1-c^2)|u|_2^2 + |u_x|_2^2) - \rho^{p+1}|u|_{p+1}^{p+1} > 0$$

for some $\rho \in (0,1)$. Hence there exists some $\rho_0 \in (0,1)$ such that $W_c(\rho_0) = R_c(\rho_0 u) = 0$. That is,

$$\rho_0^2((1-c^2)|u|_2^2 + |u_x|_2^2) = \rho_0^{p+1}|u|_{p+1}^{p+1}.$$

By Lemma 5.2, we obtain

$$\begin{aligned} d(c) &\leq L_c(\rho_0 u, -c\rho_0 u) \\ &= \rho_0^2 \left(\frac{1}{2}(1-c^2)|u|_2^2 + \frac{1}{2}|u_x|_2^2 \right) - \frac{1}{p+1}\rho_0^{p+1}|u|_{p+1}^{p+1} \\ &= \left(\frac{1}{2} - \frac{1}{p+1} \right) \rho_0^{p+1}|u|_{p+1}^{p+1} < \left(\frac{1}{2} - \frac{1}{p+1} \right) |u|_{p+1}^{p+1} \\ &= \frac{1}{2}(1-c^2)|u|_2^2 + \frac{1}{2}|u_x|_2^2 - \frac{1}{p+1}|u|_{p+1}^{p+1} \\ &\quad - \frac{1}{2}((1-c^2)|u|_2^2 + |u_x|_2^2 - |u|_{p+1}^{p+1}) \\ &= L_c(u, -cu) - \frac{1}{2}R_c(u) \leq L_c(u, v) - \frac{1}{2}R_c(u) \\ &= L_c(u_0, v_0) - \frac{1}{2}R_c(u). \end{aligned}$$

Hence $R_c(u(t)) < -2(L_c(u_0, v_0) - d(c))$ for $0 \le t < T$. This completes the proof of Lemma 5.3. □

Proof of Theorem 5.1. By the local existence theorem (Theorem 3.3), it suffices to prove the a priori estimate $\|u(t)\|_1 + |v(t)|_2 \le C(T_{\max})$ for all $0 \le t < T_{\max}$. By Lemma 5.3, we have $u(t) \in K_1^c$ for $0 \le t < T_{\max}$. This means

$$\frac{1}{2}(1-c^2)|u|_2^2 + \frac{1}{2}|u_x|_2^2 - \frac{1}{p+1}|u|_{p+1}^{p+1} < d(c)$$

and

$$(1-c^2)|u|_2^2 + |u_x|_2^2 - |u|_{p+1}^{p+1} > 0.$$

By these two inequalities, we obtain

$$(1-c^2)|u|_2^2 + |u_x|_2^2 < \frac{2(p+1)}{p-1}d(c) = \lambda^{\frac{p+3}{2(p-1)}}\|\varphi\|_1^2.$$

where $\lambda = (1-c^2)^{\frac{1}{2}}$. This implies

$$\|u(t)\|_1^2 < \lambda^{\frac{5-p}{p-1}}\|\varphi\|_1^2.$$

On the other hand, by $E(\vec{u}(t)) = E(\vec{u}_0)$, we obtain

$$\begin{aligned} |v(t)|_2^2 &\le \frac{2}{p+1}|u(t)|_{p+1}^{p+1} + 2E(\vec{u}_0) \\ &\le \frac{2}{p+1}C_{p,1}^{p+1}\|u(t)\|_1^{p+1} + 2E(\vec{u}_0) \\ &\le \frac{2}{p+1}\lambda^{\frac{(5-p)(p+1)}{2(p-1)}}\|\varphi\|_1^2 + 2E(\vec{u}_0) < +\infty \end{aligned}$$

for all $t \in [0, T_{\max})$. □

Remark. It is an open problem whether or not a solution $\vec{u}$ of the BQ initially sufficiently close to unstable solitary wave $\vec{\varphi}_c (c \neq 0)$ blows up.

Acknowledgments. This work is a portion of my thesis. I would like to thank my advisor, Professor Walter A. Strauss, for his guidance and constant encouragement. I also wish to thank Yan Guo and Kuniaki Nakamitsu for helpful discussions during the preparation of this paper, and the reviewer for helpful comments.

REFERENCES

[1] J. M. Ball, *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math. Oxford, Ser (2), 28 (1977) pp. 473–486.

[2] C. Bandle, *Isoperimetric Inequalities and Applications*, Pitman, London, 1980.

[3] M. Beals, *Self-spreading and strength of singularities for solutions to semilinear wave equations*, Ann. of Math. (2), 118 (1983) pp. 197–214.

[4] ———, *Singularities of conormal radially smooth solutions to nonlinear wave equations*, Comm. Partial Differential Equations, 13 (1988) pp. 1355–1382.

[5] H. Berestycki and T. Cazenave, *Instabilité des états stationnaires dans les équations de Schrödinger et de Klein-Gordon non-linéaires*, C. R. Acad. Sci., 293 (1981), pp. 489–492.

[6] H. Berestycki and P. Lions, *Nonlinear scalar field equations*, Arch. Rational Mech. Anal., 82 (1983), pp. 313–375.

[7] M. S. BERGER, *On the existence and structure of stationary states for a nonlinear Klein-Gordon equation,* J. Funct. Anal., 9 (1972), pp. 249–261.

[8] J. BONA AND R. SACHS, *Global existence of smooth solutions and stability of solitary waves for a generalized Boussinesq equation,* Comm. Math. Phys., 118 (1988), pp. 15–29.

[9] J. BONA AND R. SMITH, *A model for the two-way propagation of water waves in channel,* Math. Proc. Cambridge Philos. Soc., 79 (1976), pp. 167–182.

[10] J. M. BONY, *Interactions des singularities pour les equations se Klein-Gordon non lineaires,* Sem. Goulaouic-Meyer-Schwartz, exposé 10 (1983–1984).

[11] J. BOUSSINESQ, *Théorie des ondes et de remous qui se propagent,* J. Math. Pures Appl. (2), 17 (1872), pp. 55–108.

[12] L. CAFFARELLI AND A. FRIEDMAN, *Differentiability of the blow-up curve for one-dimensional nonlinear wave equations,* Arch. Rational Mech. Anal., (1987), pp. 83–98.

[13] R. E. CAFLISCH, *Shallow water waves,* unpublished manuscript.

[14] C. V. COFFMAN, *Uniqueness of the ground state solution for* $\Delta u - u + u^3 = 0$ *and a variational characterization of other solutions,* Arch. Rational Mech. Anal., 46 (1972), pp. 81–95.

[15] A. CRANNELL, *The Existence of Many Periodic Non-Travelling Solutions to the Boussinesq Equation,* Ph.D. Thesis, Brown University, Providence, RI, 1992.

[16] P. DEIFT, C. TOMEI, AND E. TRUBOWITZ, *Inverse scattering and the Boussinesq equation,* Comm. Pure Appl. Math., 35 (1982), pp. 567–628.

[17] F. FALK, E. W. LAEDKE, AND K. H. SPATSCHEK, *Stability of solitary wave pulses in shape-memory alloys,* Phys. Rev. B (3), 36 (1987), pp. 3031–3041.

[18] E. GAGLIARDO, *Proprieta di alcune classi di funzioni in piu varibili,* Ricerche Mat., 7 (1958), pp. 102–137.

[19] ———, *Ulteriori proprieta di alcune classi di funzioni in piu variabili,* Ricerche Mat., 8 (1959), pp. 24–51.

[20] R. T. GLASSEY, *Blow-up theorems for nonlinear wave equations,* Math. Z., 132 (1973), pp. 183–302.

[21] ———, *Finite-time blow-up for solutions of nonlinear wave equations,* Math. Z., 177 (1981), pp. 323–340.

[22] M. GROVES AND W. CRAIG, *Hamiltonian long-wave approximations to the water-wave problem,* to appear.

[23] B. HANOUZET AND J. L. JOLY, *Explosion pour des problemes hyperboliques semi-linéaires avec second membre noncompatible,* C. R. Acad. Sci. Paris, 301 (1985), pp. 581–584.

[24] B. HANOUZET, *Non-existence de solutions globales pour des équations des ondes non linéaires à domnéas de Cauchy petites,* C. R. Acad. Sci. Paris, 301 (1985), pp. 569–572.

[25] F. JOHN, *Blow-up of solutions of nonlinear wave equations in three space dimensions,* Manuscripta Math., 28 (1979), pp. 235–268.

[26] ———, *Blow-up for quasi-linear wave equation in three space dimensions,* Comm. Pure Appl. Math., 34 (1981), pp. 29–51.

[27] ———, *Lower bounds for the life span of solutions of nonlinear wave equations in three dimensions,* Comm. Pure Appl. Math., 36 (1983), pp. 1–36.

[28] V. KALANTAROV AND O. LADYZHENSKAYA, *The occurrence of collapse for quasilinear equations of parabolic and hyperbolic types,* J. Soviet Math., 10 (1978), pp. 53–70.

[29] T. KATO, *Blow-up of solutions of some nonlinear hyperbolic equation,* Comm. Pure Appl. Math., 33 (1980), pp. 501–505.

[30] ———, *Quasilinear equations of evolution, with applications to partial differential equations,* in Lecture Notes in Math. 448, Springer, Berlin, Heidelberg, and New York, 1974, pp. 25–70.

[31] J. B. KELLER, *On solutions of nonlinear wave equations,* Comm. Pure Appl. Math., 10 (1957), pp. 523–530.

[32] E. V. KRISHNAN, *An exact solution of the classical Boussinesq equation,* J. Phys. Soc. Japan, 51 (1982), pp. 2391–2392.

[33] M. K. KWONG, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[34] H. A. LEVINE, *Instability and nonexistence of global solutions to nonlinear wave equations of the form* $Pu_{tt} = -Au + F(u)$, Trans. Amer. Math. Soc., 192 (1974), pp. 1–21.

[35] ———, *Some additional remarks on nonexistence of global solutions to nonlinear wave equations,* SIAM J. Math. Anal., 5 (1974), pp. 138–146.

[36] ———, *An estimate for the best constant in a Sobolev inequality involving three integral norms,* Ann. Mat. Pura Appl. (4), 124 (1980), pp. 181–197.

[37] Y. Liu, *Instability of solitary waves for generalized Boussinesq equations,* J. Dynamics Differential Equations, (1993), pp. 537–558.
[38] H. P. Mckean, *Boussinesq's equation on the circle,* Comm. Pure appl. Math., 34 (1981), pp. 567–628.
[39] K. Mcleod and J. Serrin, *Uniqueness of solutions of semilinear Poisson equations,* Proc. Nat. Acad. Sci. U.S.A. 78 (1981), pp. 6592–6595.
[40] R. Melrose and N. Ritter, *Interaction of nonlinear progressing waves,* Ann. of Math. (2), 121 (1985), pp. 187–213.
[41] B. V. Sz. Nagy, *Über integralgleichungen zwischen einer Funktion und ihrer Ableitung,* Acta Sci. Math. (Szeged), 10 (1941), pp. 64–74.
[42] Z. Nehari, *On a nonlinear differential equation arising in nuclear physics,* Proc. Roy. Irish Acad. Sci. Sect. A, 62 (1963), pp. 117–135.
[43] L. Nirenberg, *Remarks on strongly elliptic partial differential equations,* Comm. Pure Appl. Math., 8 (1955), pp. 648–674.
[44] R. Osserman, *The isoperimetric inequality,* Bull. Amer. Math. Soc., 84 (1978), pp. 1182–1238.
[45] L. E. Payne, *Uniqueness criteria for steady state solutions of Navier Stokes equations,* Simpos. Internoz. Appl. Anal. Fix. Mat. (Cagliari-Sas-sari, 1964) Edizioni Cremonese, Roma, 1965, pp. 130–153.
[46] L. E. Payne and D. H. Sattinger, *Saddle points and instability of nonlinear hyperbolic equations,* Israel J. Math., 22 (1975), pp. 273–303.
[47] R. L. Pego and M. I. Weinstein, *Eigenvalues, and instability of solitary waves,* Philos. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.
[48] J. Rauch and M. Reed, *Singularities produced by the nonlinear interaction of three progressing waves: Examples,* Comm. Partial Differential Equations (1982), pp. 1117–1133.
[49] G. H. Ryder, *Boundary value problems for a class of nonlinear differential equations,* Pacific J. Math., 22 (1967), pp. 477–503.
[50] J. Schaeffer, *The equations* $u_{tt} - \Delta u = |u|^p$ *for the critical value of* p, Proc. Roy. Soc. Edinburgh, Sect. A, 101 (1985).
[51] J. Shatah, *Weak solutions and development of singularities of the* SU(2) σ*-model,* Comm. Pure Appl. Math., 41 (1988), pp. 459–469.
[52] T. Sideris, *Nonexistence of global solutions to semilinear wave equations in high dimensions,* J. Differential Equations, 52 (1984), pp. 378–406.
[53] B. Straughan, *Further nonexistence theorems for abstract nonlinear wave equations,* Proc. Amer. Math. Soc., 48 (1975), pp. 381–390.
[54] W. A. Strauss, *Existence of solitary waves in higher dimensions,* Comm. Math. Phys., 55 (1977), pp. 149–162.
[55] J. L. Synge, *On a certain non-linear differential equation,* Proc. Roy. Irish Acad. Sci. Sect. A, 62 (1961–1962), pp. 17–41.
[56] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates,* Comm. Math. Phys., 87 (1983), pp. 567–576.
[57] V. E. Zakharov, *On the problem of stochastization of one-dimensional chains of nonlinear operators,* Zh. Eksper. Teoret. Fiz., 65 (1973), pp. 219–228.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1547–1563, November 1995

010

SMOOTHING PROPERTIES, DECAY, AND GLOBAL EXISTENCE OF SOLUTIONS TO NONLINEAR COUPLED SYSTEMS OF THERMOELASTIC TYPE*

JAIME E. MUÑOZ RIVERA† AND REINHARD RACKE‡

Abstract. We consider a nonlinear coupled system of evolution equations, the simplest of which models a thermoelastic plate. Smoothing and decay properties of solutions are investigated as well as the local well-posedness and the global existence of solutions. For the system of standard thermoelasticity it is proved that there is no similar smoothing effect.

AMS subject classifications. 35K22, 73B30

Key words. thermoelasticity, thermoelastic plates, smoothing, exponential and polynomial decay, global solution, initial boundary value problems

1. Introduction. In this paper we consider regularizing properties of systems that are regarded as models for thermoelastic plate equations. We will show that the vertical deflection of the plate as well as the temperature are arbitrarily smooth for positive times, no matter what regularity the initial vertical deflection and the initial temperature have. We will show this fact in §3. This property is not valid for other thermoelastic models such as the thermoelastic bar, for example, as we shall see in §4. More generally, we consider a nonlinear coupled thermoelastic plate modelled in a separable Hilbert space $\mathcal{H}$ by

$$u_{tt} + M([u,\theta])A^2u + N([u,\theta])(A+\mu)\theta = 0, \tag{1.1}$$

$$\theta_t + R([u,\theta])(A+\alpha)\theta - Q([u,\theta])(A+\mu)u_t = 0. \tag{1.2}$$

Here M, N, R, $Q : \mathbb{R}^5 \longrightarrow \mathbb{R}$ are C^2-functions, and M, R, and NQ are strictly positive; $\alpha, \mu \in \mathbb{R}$. Finally, by $[u,\theta]$ we denote the following vector field:

$$[u,\theta](t) := (||u_t||^2, ||A^{\frac{1}{2}}u||^2, ||A^{\frac{1}{2}}u_t||^2, ||Au||^2, ||A^{\frac{1}{2}}\theta||^2)(t),$$

where $||\cdot||$ denotes the norm in $\mathcal{H}$; $A : D(A) \subset \mathcal{H} \longrightarrow \mathcal{H}$ is a nonnegative, self-adjoint operator. The solution (u,θ) will satisfy the initial conditions

$$u(t=0) = u_0, \quad u_t(t=0) = u_1, \quad \theta(t=0) = \theta_0, \tag{1.3}$$

and the abstract "boundary" conditions

$$u(t) \in D(A^2), \quad \theta(t) \in D(A), \quad t \geq 0. \tag{1.4}$$

* Received by the editors June 15, 1993; accepted for publication (in revised form) January 31,1994.

† Department of Research and Development, National Laboratory for Scientific Computation, Rua Lauro Müller 455, 22290 Rio de Janeiro, RJ, Brazil and Instituto de Matemática, Federal University of Rio de Janeiro. The research of this author was supported by a Conselho Nacional de Desenvolvimento Científico e Tecnológico–Gesellschaft für Mathematik und Datenverarbeitung grant.

‡ Fakultät für Mathematik, Universität Konstanz, Universitätsstraße 10, 78434 Konstanz 1, Germany. The research of this author was supported by the Sonderforschungsbereich 256 at the University of Bonn.

A simple example is the following system modelling a thermoelastic plate in the linearized version;

$$u_{tt} + \Delta^2 u + \beta \Delta \theta = 0 \quad \text{in} \quad [0, \infty[\times \Omega, \tag{1.5}$$

$$\theta_t - \Delta \theta - \beta \Delta u_t = 0 \quad \text{in} \quad [0, \infty[\times \Omega, \tag{1.6}$$

where $\beta \neq 0$. The boundary $\partial\Omega$ of the open set Ω is assumed to be smooth; u and θ will satisfy

$$u = \Delta u = 0 \quad \text{on} \quad \partial\Omega, \quad \theta = 0 \quad \text{on} \quad \partial\Omega. \tag{1.7}$$

Kim [6] studied equations (1.5), (1.6) in a bounded domain with the boundary condition (1.7) for u replaced by $u \in H_0^2(\Omega)$, and showed exponential decay of the couple (u, θ).

We are first interested in proving smoothing properties, i.e., the solution (u, θ) is arbitrarily smooth for $t > 0$ no matter which regularity the initial data have. Smoothness for the abstract system (1.1)–(1.4) means that the solution $(u(t), \theta(t))$ belongs to $D(A^m) \times D(A^m)$ for any $m \in \mathbb{N}$ and any $t > 0$. Then we shall investigate the rate of decay for the couple (u, θ) as $t \to +\infty$, depending on A, and in case (1.5)–(1.7) naturally depending on the domain Ω. Finally we show the global existence of solutions (u, θ) if A is strictly positive. These results describe system (1.1)–(1.2) as parabolic, the similarities to solutions of heat equations will be obvious. In contrast to this we study the system of standard thermoelasticity (cf. [15], [16]), which is written as follows in the simplest one-dimensional case (thermoelastic bar equation):

$$u_{tt} - \tau u_{xx} + \gamma \theta_x = 0, \tag{1.8}$$

$$\theta_t - \kappa \theta_{xx} + \gamma u_{xt} = 0, \tag{1.9}$$

$$u(t = 0) = u_0, \quad u_t(t = 0) = u_1, \quad \theta(t = 0) = \theta_0, \tag{1.10}$$

$$u = \theta_x = 0 \quad \text{on} \quad \partial\Omega, \tag{1.11}$$

where $\Omega =]0, 1[$, $\Omega =]0, +\infty[$, or $\Omega =]-\infty, +\infty[$; (u, θ) is a function of $t \geq 0$ $x \in \Omega$, and τ, $|\gamma|$, κ are positive constants. It is known (cf. [16]) that solutions behave like solutions to the heat equation with respect to the decay behavior; but it is not true for n-dimensional thermoelastic systems if $n \geq 2$. It is well known by now that, in this case for the whole space $\mathbb{R}^n$, the displacement vector field can be decomposed in two parts: the solenoidal part, which satisfies the wave equation, and the irrotational part which is a gradient (see [11]). Clearly, the solenoidal part propagates singularities. We shall prove that the smoothing property does not hold even for the irrotational part. Moreover, we shall prove, that it behaves like a wave equation, which propagates singularities. For the formulation of the result we introduce the following notation. Ω will denote a domain in $\mathbb{R}^n$.

$$H^m(\Omega) = W^{m,2}(\Omega), \quad H_0^m(\Omega) = W_0^{m,2}(\Omega); \quad m \in \mathbb{N},$$

will denote the usual Sobolev spaces based on $L^2(\Omega)$ (cf. [1]); ∇ will denote the gradient, $\langle \cdot, \cdot \rangle$ will denote the inner product in $L^2(\Omega)$ or in a general separable Hilbert

space $\mathcal{H}$, $|\cdot|$ will denote the norm in $L^2(\Omega)$, $C^k(I,E)$, $k \in \mathbb{N}$ will denote the space of k-times continuously differentiable functions from $I \subset \mathbb{R}$ into a Banach space E, analogously, $L^p(I,E)$, $1 \le p \le \infty$.

The smoothing properties for the systems (1.1), (1.2) and (1.5), (1.6), respectively, are expressed in Theorem 3.1. The local existence of solutions is the subject of Theorem 2.4. To describe the decay, we consider the linearized version of (1.1), (1.2) assuming $\alpha = \mu = 0$ (only for simplicity; in the general case α is nonnegative and μ is such that the product $\mu q(t)$ is positive), i.e.,

$$u_{tt} + m(t)A^2u + n(t)A\theta = 0, \tag{1.12}$$

$$\theta_t + r(t)A\theta - q(t)Au_t = 0, \tag{1.13}$$

where m, n, r, q are C^1-functions of t satisfying

$$m_0 \le m(t) \le m_1, \quad n_0 \le |n(t)| \le n_1, \quad q_0 \le |q(t)| \le q_1, \quad r_0 \le r(t) \le r_1$$

with $m_0, \dots, r_1$ being positive real numbers, $n(t)q(t) > 0$ $\forall t > 0$, and similar bounds for the derivatives of them m', n', r', q'.

The asymptotic behavior of the solution depends on the spectral properties of the operator A. When A is coercive, we get exponential decay (see Theorem 3.5), which implies, in particular, the decay of solutions for the thermoelastic plates given by (1.5)–(1.7) when Ω is a bounded domain. If the spectrum of A approaches zero, one needs more information on A than that given in the general setting. In Theorem 3.6 we present a typical result that has L^2- and L^∞-decay rates for the thermoelastic plate equation (1.5), (1.6) if Ω is the whole space $\mathbb{R}^n$ or if Ω is an exterior domain. By interpolation one also gets decay rates in $L^q(\Omega)$ for $2 < q < \infty$.

For the case $A \ge \nu > 0$, $\alpha = \mu = 0$ we shall extend our local existence result to a global existence result (see Theorem 2.8).

We remark that right-hand sides in (1.1) and (1.2), respectively, with appropriate regularity (for Theorem 2.4) and smallness (for Theorem 2.8) can easily be included.

Using Theorem 3.6 it would also be possible to prove a global existence result for (1.5)–(1.7), (1.3) for small data in exterior domains, including the whole space $\mathbb{R}^n$ (cf. [17]); we do not go into details here. Finally, we turn to system (1.8)–(1.11) in standard thermoelasticity and related systems as (1.8)–(1.10) with the boundary conditions

$$u_x = \theta = 0 \quad \text{on} \quad \partial\Omega, \tag{1.14}$$

or systems in higher dimensions of the following type:

$$u_{tt} - \tau\Delta u + \gamma\nabla\theta = 0, \tag{1.15}$$

$$\theta_t - \kappa\Delta\theta + \gamma \operatorname{div} u_t = 0, \tag{1.16}$$

$$u(t) \in \overline{\nabla H_0^1(\Omega)} \quad \forall t \ge 0 \tag{1.17}$$

for $\Omega = \mathbb{R}^2$ or $\Omega = \mathbb{R}^3$. For domains in $\mathbb{R}^3$ with smooth boundary we will consider boundary conditions of the form

$$\operatorname{div} u = \theta = 0 \quad \text{on} \quad \partial\Omega. \tag{1.18}$$

(According to (1.10), the initial condition has to be satisfied in each case.)

Systems (1.8)–(1.10), (1.11) and (1.8)–(1.10), (1.14), respectively, describe the initial boundary value problem for a one-dimensional thermoelastic bar with rigidly clamped and thermally insulated boundary in the case of (1.11), and with traction free boundary at constant temperature in the case of (1.14). System (1.15)–(1.17) describes the dissipative part of the solution to the Cauchy problem in $\mathbb{R}^2$ or $\mathbb{R}^3$; cf. [12], [15]. System (1.15)–(1.18) is a set of equations for the dissipative part of a thermoelastic problem in $\mathbb{R}^3$ with the specific boundary condition given abovep; cf. [13].

In each case the solution (u, θ) has the same decay rates as the solution to the heat equation (see [3], [5], [10]–[13], [15], [18], [19]; for a survey cf. [16]). In contrast to this we shall prove that they do not have the same smoothing property: singularities in the initial data are propagated as time increases. This shows that the coupling for thermoelastic plates is much stronger than that in standard thermoelasticity. All problems above can be considered simultaneously, namely, for (1.8), (1.9), or (1.15), (1.16) with boundary conditions given by (1.17), (1.18), it is easy to see that $v := u$ satisfies

$$v_{ttt} - \kappa\Delta v_{tt} - (\gamma^2 + \tau)\Delta v_t + \kappa\tau\Delta^2 v = 0, \tag{1.19}$$

(1.20)
$$v(t=0) = v_0 := u_0, \quad v_t(t=0) = v_1 := u_1, \quad v_{tt}(t=0) = v_2 := \tau\Delta u_0 - \gamma\nabla\theta_0,$$

as well as

$$v = \Delta v = 0 \quad \text{on} \quad \partial\Omega \tag{1.21}$$

in case (1.11), and

$$v_x = v_{xxx} = 0 \quad \text{on} \quad \partial\Omega \tag{1.22}$$

in case (1.14). For (1.10), (1.15)–(1.18) $v := \operatorname{div} u$ also satisfies (1.19), as well as (1.20) with

$$v_0 = \operatorname{div} u_0, \quad v_1 := \operatorname{div} u_1, \quad v_2 := \tau\Delta\operatorname{div} u_0 - \gamma\Delta\theta_0,$$

and the boundary condition (1.21). We remark that θ satisfies a differential equation similar to v with appropriate boundary conditions. Denoting by A the Laplace operator with domain

$$D(A) := H^2(\Omega) \cap H_0^1(\Omega), \quad Av := -\Delta v$$

(respectively,

$$D(A) := \left\{v \in H^2(\Omega); \forall\varphi \in H^1(\Omega); \langle\nabla v, \nabla\varphi\rangle = -\langle v, \Delta\varphi\rangle\right\}; \quad Av := -\Delta v),$$

we see that v satisfies

$$v_{ttt} + \kappa A v_{tt} + (\gamma^2 + \tau)A v_t + \kappa\tau A^2 v = 0, \tag{1.23}$$

$$v(t=0) = v_0, \quad v_t(t=0) = v_1, \quad v_{tt}(t=0) = v_2, \tag{1.24}$$

$$v(t) \in D(A^2), \quad t \geq 0. \tag{1.25}$$

(1.23)–(1.25) will be considered in a separable Hilbert space $\mathcal{H}$ again, $v : [0, \infty[\to \mathcal{H}$, and our result on propagation of singularities will be proved in Theorem 4.1.

2. Existence results. First we study the linearized problem

$$u_{tt} + m(t)A^2u + n(t)(A+\mu)\theta = f_1, \tag{2.1}$$

$$\theta_t + r(t)(A+\alpha)\theta - q(t)(A+\mu)u_t = f_2, \tag{2.2}$$

$$u(t=0) = u_0, \quad u_t(t=0) = u_1, \quad \theta(t=0) = \theta_0 \tag{2.3}$$

(with more general right-hand sides), and we look for solutions (u, θ) satisfying

$$u \in C^2([0,\infty[, \mathcal{H}) \cap C^1([0,\infty[, D(A)) \cap C^0([0,\infty[, D(A^2)), \tag{2.4}$$

$$\theta \in C^1([0,\infty[, \mathcal{H}) \cap C^0([0,\infty[, D(A)). \tag{2.5}$$

Rewriting (2.1)–(2.3) as a first-order system for

$$V := \begin{pmatrix} u_t \\ Au \\ \theta \end{pmatrix}, \quad V^0 := \begin{pmatrix} u_1 \\ Au_0 \\ \theta_0, \end{pmatrix},$$

we consider

$$V_t + \underbrace{\begin{pmatrix} 0 & m(t)A & n(t)(A+\mu) \\ -A & 0 & 0 \\ -q(t)(A+\mu) & 0 & r(t)(A+\alpha) \end{pmatrix}}_{=:B(t)} V(t) = \underbrace{\begin{pmatrix} f_1 \\ 0 \\ f_2 \end{pmatrix}}_{=:F(t)}, \tag{2.6}$$

$$V(t=0) = V^0. \tag{2.7}$$

The coefficients m, n, q, r are considered C^1-functions. For $V = (V^1, V^2, V^3)$, $W = (W^1, W^2, W^3)$, $t \geq 0$, the Hilbert space $\mathcal{H}_t$ is defined by the inner product

$$\langle V, W\rangle_t := \langle V^1, W^1\rangle_{\mathcal{H}} + m(t)\langle V^2, W^2\rangle_{\mathcal{H}} + \frac{n(t)}{q(t)}\langle V^3, W^3\rangle_{\mathcal{H}}.$$

Observe that the corresponding norm $||\cdot||_t$ is equivalent to the norm $||\cdot||$ in $\mathcal{H}$. Defining the operator $B(t)$ by $D(B(t)) := D(A)$ (in each component), it is not difficult to see that $-B(t)$ generates a C_0-semigroup with constants $M = 1$, $\hat{\beta} = \max\{0, -\alpha\}$, and hence $\{B(t)\}_{t\geq 0}$ is a stable family of negative generators in $\mathcal{X} = H$ with stability constants (M, β), β depending on m', n', q'. With $\mathcal{Y} := D(A)$ we see that

$$||B(t)v - B(s)v||_{\mathcal{X}} \leq |t-s|\,||v||_{\mathcal{Y}} \quad \text{for} \quad 0 \leq t,\ s \leq T,\ v \in \mathcal{Y}.$$

Therefore $(\{B(t)\}_{t\geq 0}, \mathcal{X}, \mathcal{Y})$ are a CD-system in the terminology of Kato [4]. As a consequence we have the following lemma.

LEMMA 2.1. *For $V^0 \in \mathcal{Y}$, $F : [0,T] \to \mathcal{X}$ Lipschitz continuous, there is a unique solution $V \in C^1([0,T], \mathcal{X}) \cap C^0([0,T], \mathcal{Y})$ of* (2.6), (2.7).

COROLLARY 2.2. *For $u_0 \in D(A^2)$, $u_1 \in D(A)$, $\theta_0 \in D(A)$, $f_1, f_2 : [0,T] \to \mathcal{X}$ Lipschitz continuous, there is a unique solution (u, θ) of* (2.1)–(2.3) *satisfying* (2.4), (2.5).

The higher regularity for more regular data is given by the following lemma, where we assume that $f_1 = 0, f_2 = 0$ for simplicity.

LEMMA 2.3. *Let $k \geq 2$, $u_0 \in D(A^k)$, $u_1 \in D(A^{k-1})$, $\theta_0 \in D(A^{k-1}), m,\ n,\ r,\ q,\ \in C^{k-1}$. Then there is a unique (u, θ) to (2.1)–(2.3) satisfying*

$$u \in \cap_{j=0}^{k} C^j([0,\infty[, D(A^{k-j})), \quad \theta \in \cap_{j=0}^{1} C^j([0,\infty[, D(A^{k-j-1})).$$

Proof of Lemma 2.3. $\exists \lambda \geq 0$, $(A+\lambda)^{-1} : \mathcal{H} \to \mathcal{H}$ is bounded; let $w := (A+\lambda)u$, $\psi := (A+\lambda)\theta$. Formally, we obtain, assuming $\lambda = 0$ for simplicity,

$$w_{tt} + m(t)A^2 w + n(t)(A+\mu)\psi = 0, \tag{2.8}$$

$$\psi_t + r(t)(A+\alpha)\psi - q(t)(A+\mu)w_t = 0, \tag{2.9}$$

$$w(t=0) = Au_0, \quad w_t(t=0) = Au_1, \quad \psi(t=0) = A\theta_0. \tag{2.10}$$

Observe that $||A^{-1}w|| \leq c||w||$ holds. Assuming $u_0 \in D(A^3)$, $u_1 \in D(A^2), \theta_0 \in D(A^2)$, we can solve (2.8)–(2.10) with Corollary 2.2. Then $\hat{u} =: A^{-1}w, \hat{\theta} =: A^{-1}\psi$ solves (2.1)–(2.3), and hence it is equal to u and θ, respectively. From the regularity of (w, ψ) we conclude

$$u \in \cap_{j=0}^{3} C^j([0,\infty[, D(A^{3-j})), \quad \theta \in \cap_{j=0}^{2} C^j([0,\infty[, D(A^{2-j})),$$

where we used the differential equations and needed $m,\ n,\ q,\ r$ to be C^1-functions. The case $k \geq 4$ is obtained taking $w(t=0) = A^k u_0$, $w_t(t=0) = A^k u_1$, $\psi(t=0) = A^k \theta_0$. □

Our local existence result is summarized in the following theorem.

THEOREM 2.4. *Let $k \geq 3$, let $M,\ N,\ Q,\ R \in C^{k-1}(\mathbb{R}^5, \mathbb{R})$ with M, R, and the product NQ being positive functions, and let*

$$(u_0, u_1, \theta_0) \in D(A^k) \times D(A^{k-1}) \times D(A^{k-1}).$$

Then there exists a unique solution (u, θ) to (1.1)–(1.4) satisfying

$$(u, \theta) \in \cap_{j=0}^{k} C^j([0,T], D(A^{k-j})) \times \cap_{j=0}^{k-1} C^j([0,T], D(A^{k-1-j}))$$

for some $T > 0$. T depends only on the initial data $T = T(\rho)$, where

$$\rho := (||u_0||_{D(A^2)}, ||u_1||_{D(A)}, ||\theta_0||_{D(A)})$$

and $T \to \infty$ as $\rho \to 0$.

To prove Theorem 2.4 we shall use a fixed point argument in appropriate spaces. Let $u_0 \in D(A^3)$, $u_1 \in D(A^2)$, $\theta_0 \in D(A^2)$. For $N_1 > 0$, $N_2 > 0$, $T > 0$ let

$$\begin{aligned}
&\mathcal{X}(N_1, N_2, T) \\
&\quad =: \Big\{ (u,\theta) : [0,T] \to \mathcal{H}; \partial_t^j u \in L^\infty([0,T]; D(A^{3-j})), j = 1,2,3.\ \partial_t^k \theta \\
&\qquad \in L^\infty([0,T]; D(A^{2-k}));\ k = 0,1,2 \Big\}
\end{aligned}$$

intersect with the set of couples (u,θ) satisfying

$$u(t=0)=u_0, \quad u_t(t=0)=u_1, \quad \theta(t=0)=\theta_0,$$

$$\sup_{0\le t\le T}\left\{\sum_{j=0}^{2}||\partial_t^j u(t)||^2_{D(A^{2-j})}+\sum_{k=0}^{1}||\partial_t^k\theta(t)||^2_{D(A^{1-k})}\right\}\le N_1^2,$$

$$\sup_{0\le t\le T}\left\{\sum_{j=0}^{3}||\partial_t^j u(t)||^2_{D(A^{3-j})}+\sum_{k=0}^{2}||\partial_t^k\theta(t)||^2_{D(A^{2-k})}\right\}\le N_2^2.$$

We observe that $\mathcal{X}(N_1,N_2,T)\neq\emptyset$ if $N_j=N_j(||u_0||_{D(A^{1+j})},||u_1||_{D(A^j)},||\theta_0||_{D(A^j)})$, $j=1,2$ is large enough.

LEMMA 2.5. *$\mathcal{X}(N_1,N_2,T)$ is a closed subspace of the complete metric space $\mathcal{Z}$ defined by*

$$\mathcal{Z}:=\left\{(u,\theta):[0,T]\to\mathcal{H};u_t,\ \theta\in L^\infty([0,T],D(A^{\frac12})),\ u\in L^\infty([0,T],D(A))\right\}$$

and the metric

$$d((u,\theta),(v,\eta)):=||(u_t-v_t,A^{\frac12}(u_t-v_t),u-v,Au-Av,\theta-\eta,A^{\frac12}\theta-A^{\frac12}\eta)||_{L^\infty([0,T],\mathcal{H})}.$$

The standard proof (cf. [18]) of Lemma 2.5 exploits the weak–$*$ compactness of bounded sets in $L^\infty([0,T],\mathcal{H})$ (observe that $\mathcal{H}$ is assumed to be separable).

A mapping $\mathcal{S}:\mathcal{X}(N_1,N_2,T)\subset\mathcal{Z}\to\mathcal{Z}$ is defined by $\mathcal{S}(u,\theta):=(\hat u,\hat\theta):=$ solution to

$$\hat u_{tt}+M([u,\theta])A^2\hat u+N([u,\theta])(A+\mu)\hat\theta=0,$$

$$\hat\theta_t+R([u,\theta])(A+\alpha)\hat\theta-Q([u,\theta])(A+\mu)\hat u_t=0,$$

$$\hat u(t=0)=u_0, \quad \hat u_t(t=0)=u_1, \quad \hat\theta(t=0)=\theta_0,$$

which exists according to Lemma 2.3. Observe that $\frac{d}{dt}M([u,\theta](t))$ is bounded since $(u,\theta)\in\mathcal{X}(N_1,N_2,T)$.

LEMMA 2.6. *The mapping $\mathcal{S}$ defined above maps $\mathcal{X}$ into itself if T is sufficiently small depending on N_1.*

Proof of Lemma 2.6. Let us denote

$$\hat V=\begin{pmatrix}\hat u_t\\ A\hat u\\ \hat\theta\end{pmatrix}, \quad \hat V_0=\begin{pmatrix}\hat u_1\\ A\hat u_0\\ \hat\theta_0\end{pmatrix}.$$

Then $\hat V$ satisfies

$$\hat V_t(t)+\hat B(t)\hat V(t)=0, \quad \hat V(t=0)=V_0,$$

where $\hat{B}$ equals the previously defined $B(t)$ with $m(t) := M([u,\theta](t))$ and so on. We have that

$$||\hat{V}(t)|| \leq Me^{\beta t}||\hat{V}_0||$$

with $\beta \leq cN_1^2$ (cf. [4]), since β depends on m', n', q' essentially and

$$||A\hat{V}(t)|| + ||\hat{V}_t(t)|| \leq \hat{M}e^{\hat{\beta}t}||\hat{V}_0||_{D(A)}$$

with $\hat{\beta} \leq cN_1^2$. $\hat{W} := A\hat{V}$ satisfies

$$\hat{W}_t(t) + \hat{B}(t)\hat{W}(t) = 0, \quad \hat{W}(t=0) = A\hat{V}^0,$$

and hence we also obtain

$$||A\hat{V}_t(t)|| + ||A^2\hat{V}(t)|| \leq \hat{M}e^{\hat{\beta}t}||A\hat{V}^0||_{D(A)}.$$

This implies $(u,\theta) \in \mathcal{X}(N_1, N_2, T)$ if

$$N_1^2 \geq 4||\hat{V}^0||^2_{D(A)}, \quad N_2^2 \geq 4||\hat{V}^0||^2_{D(A)}, \quad Me^{\beta T} \leq 2,$$

which is true if

$$T \leq \frac{1}{2cN_1^2}\log\left(\frac{2}{N_1}\right). \qquad \square \tag{2.11}$$

LEMMA 2.7. *The mapping $\mathcal{S}$ defined above is a contraction mapping if T is sufficiently small depending on N_1.*

Proof of Lemma 2.7. Let $(\hat{u}^j, \hat{\theta}^j) := \mathcal{S}(u^j, \theta^j)$ $j = 1,2$ and let (2.11) be satisfied. Then $w := \hat{u}^1 - \hat{u}^2$, $\psi := \hat{\theta}^1 - \hat{\theta}^2$ satisfy

$$w_{tt} + m(t)A^2w + n(t)(A+\mu)\psi = \hat{m}(t)A^2\hat{u}^2 + \hat{n}(t)(A+\mu)\hat{\theta}^2,$$

$$\psi_t + r(t)(A+\mu)\psi - q(t)(A+\mu)w_t = \hat{r}(t)(A+\mu)\hat{\theta}^2 - \hat{q}(t)(A+\mu)\hat{u}_t^2,$$

$$w(t=0) = 0, \quad w_t(t=0) = 0, \quad \psi(t=0) = 0,$$

where $m := M([u^1,\theta^1])$, $\hat{m} := M([u^1,\theta^1]) - M([u^2,\theta^2])$, and so on. Let (cf. (3.9)) $K_3 = K_3(w,\psi)$ be given by

(2.12)

$$K_3 := \frac{1}{2}\left\{||w_t||^2 + m||Aw||^2 + \frac{n}{q}||\psi||^2 + \frac{\varepsilon q_0}{4}\langle w_t, Aw\rangle - \varepsilon\langle\psi, w_t\rangle + ||Aw_t||^2 \right.$$
$$\left. + m(t)||A^2w||^2 + \frac{n}{q}||A\psi||^2 + \frac{\varepsilon q_0}{4}\langle Aw_t, A^2w\rangle - \varepsilon\langle A\psi, Aw_t\rangle\right\}.$$

Using multiplicative techniques we obtain for sufficiently small ε,

$$\frac{d}{dt}K_3(t) \leq c_1(N_1^2)K_3(t) + c_2N_1^2\left\{|\hat{m}(t)|^2 + |\hat{n}(t)|^2 + |\hat{r}(t)|^2 + |\hat{q}(t)|^2\right\}$$
$$\leq c_1(N_1^2)K_3(t) + c_3(N_1^2)d^2((u^1,\theta^1),(u^2,\theta^2)),$$

where $c_j \geq 0$, $j = 1, 2, 3$. This implies

$$\sup_{0 \leq t \leq T} K_3(t) \leq c_4(N_1^2) N_1^2 T d^2((u^1, \theta^1), (u^2, \theta^2))$$

if ε is small enough. We obtain

$$\begin{aligned} d^2((\hat{u}^1, \hat{\theta}^1), (\hat{u}^2, \hat{\theta}^2)) &\leq c_5(N_1^2) N_1^2 T d^2((u^1, \theta^1), (u^2, \theta^2)) \\ &\leq \sigma^2 d^2((u^1, \theta^1), (u^2, \theta^2)) \end{aligned}$$

with $0 < \sigma < 1$ if $T = T(N_1)$ is small enough.

The unique fixed point of $\mathcal{S}$ in $\mathcal{X}(N_1, N_2, T)$ is the desired solution (u, θ) in Theorem 2.4 for $k = 3$. The case $k \geq 4$ can be dealt with either by studying a corresponding $\mathcal{X}(N_1, N_2, \ldots, N_{k-1}, T)$ or inclusively proving the higher regularity of (u, θ) by introducing $w := A^{-1}u$, $\psi := A^{-1}\theta$ as in the proof of Lemma 2.3. (Observe that the new nonlinearities look like $\hat{M}([w, \psi]) = M([A^{-1}w, A^{-1}\psi])$ and that they are easier to deal with since they are of lower order.) This completes the proof of Theorem 2.4. □

THEOREM 2.8. *Let $A \geq \nu > 0$, $\alpha = \mu = 0$, and $k \geq 3$. Then there is $\delta > 0$ with the following property: For any*

$$(u_0, u_1, \theta_0) \in D(A^k) \times D(A^{k-1}) \times D(A^{k-1})$$

satisfying

$$||u_0||_{D(A^2)} + ||u_1||_{D(A)} + ||\theta_0||_{D(A)} < \delta,$$

there exists a unique global solution (u, θ) of (1.1)–(1.4) satisfying

$$(u, \theta) \in \cap_{j=0}^{k} C^j([0, \infty[, D(A^{k-j})) \times \cap_{j=0}^{k-1} C^j([0, \infty[, D(A^{k-1-j})).$$

Moreover, (u, θ) decays exponentially.

Proof. Let (u, θ) be a local solution according to Theorem 2.4. Under the assumption of Theorem 2.8 we obtain for

$$K_4 := K_3(u, \theta),$$

K_3 having been defined in (2.12),

$$\frac{d}{dt} K_4(t) \leq c_1(N_1^2) K_4^2(t) - d_1(N_1^2) K_4(t)$$

with c_1, $d_1 > 0$ depending on N_1^2. If $N_1^2 \leq 1$ we have

$$c_1(N_1^2) \leq c_0, \quad d_1(N_1^2) \geq d_0 > 0$$

with c_0, d_0 being independent of N_1 and t. Then

$$\frac{d}{dt} K_4(t) \leq c_0 K_4^2(t) - d_0 K_4(t).$$

By standard arguments, using Gronwall's inequality, this implies that if $K_4(0)$ is sufficiently small, then

$$K_4(s) \leq e^{-\frac{d_0}{2} s} K_4(0) \tag{2.13}$$

holds on some interval $0 \leq s \leq t_1 > 0$. This yields an a priori estimate in $s = t_1$ and by a continuation argument the solution exists globally and satisfies (2.13) for all $s \in \mathbb{R}$. The smallness of $K_4(0)$ is guaranteed by choosing $||u_0||_{D(A^2)} + ||u_1||_{D(A)} + ||\theta_0||_{D(A)}$ small enough. □

3. Smoothing effect. The main result of this section is given by the following theorem.

THEOREM 3.1. *Let* $(u,\theta) \in \cap_{j=0}^{2} C^j([0,T], D(A^{2-j})) \times \cap_{j=0}^{1} C^j([0,T], D(A^{1-j}))$ *be a solution of* (1.1)–(1.4) *for some* $T > 0$ *with* $(u_0, u_1, \theta_0) \in D(A^2) \times D(A) \times D(A)$. *Then for any* $t \in]0,T]$ *and all* $m \in \mathbb{N}$, *we have that* $(u(t), \theta(t)) \in D(A^m) \times D(A^m)$.

Proof. Let us denote $m(t) := M([u,\theta](t))$, $n(t) := N([u,\theta](t))$, $r(t) := R([u,\theta](t))$, $q(t) := Q([u,\theta](t))$. Then $m,\ n,\ r,\ q \in C^1([0,T])$ since

$$\frac{d}{dt}|[u,\theta]| \le c\left\{||u_{tt}||^2 + ||u_t||^2 + ||Au_t||^2 + ||Au||^2 + ||\theta||^2 + ||\theta_t||^2 + ||A\theta||^2\right\}.$$

By the spectral theorem for self-adjoint operators (cf. [2], [8]) there exists a Hilbert space

$$\tilde{\mathcal{H}} = \int_{\oplus} \mathcal{H}(\lambda) d\mu(\lambda),$$

a direct integral of Hilbert spaces $\mathcal{H}(\lambda)$, $\lambda \in \mathbb{R}$ with respect to a pointwise measure μ, and a unitary operator $\mathcal{U} : \mathcal{H} \to \tilde{\mathcal{H}}$ such that

$$D(A^m) = \left\{v \in \mathcal{H}; \lambda \mapsto \lambda^m \mathcal{U}v(\lambda) \in \tilde{\mathcal{H}}\right\}, \quad m \in \mathbb{N}_0$$

and

$$\mathcal{U}(A^m v)(\lambda) = \lambda^m \mathcal{U}v(\lambda).$$

Moreover,

$$||A^m v||^2 = \int_0^\infty \lambda^{2m} |\mathcal{U}v(\lambda)|^2 \, d\mu(\lambda).$$

Let us denote $v := \mathcal{U}u$, $\psi := \mathcal{U}\theta$. Then (1.1), (1.2) turn into

$$v_{tt} + m(t)\lambda^2 v + n(t)(\lambda + \mu)\psi = 0, \tag{3.1}$$

$$\psi_t + r(t)(\lambda + \alpha)\psi - q(t)(\lambda + \mu)v_t = 0, \tag{3.2}$$

where we have dropped the parameters t and λ in v and ψ. Let

$$\mathcal{E}(t,\lambda) := \frac{1}{2}|v_t(t,\lambda)|^2 + \frac{m(t)}{2}\lambda^2 |v(t,\lambda)|^2 + \frac{n(t)}{2q(t)}|\psi(t,\lambda)|^2,$$

where $|\cdot|$ is understood to be in $\mathcal{H}(\lambda)$. Multiplying equation (3.1) by v_t and (3.2) by $\frac{n}{q}\psi$ and summing up, we get

$$\frac{d}{dt}\mathcal{E}(t,\lambda) = -\frac{n(t)r(t)}{q(t)}(\lambda + \alpha)|\psi|^2 + \frac{m'(t)}{2}\lambda^2 |v|^2 + \frac{d}{dt}\left\{\frac{n}{2q}\right\}|\psi|^2, \tag{3.3}$$

$$\frac{d}{dt}\mathrm{Re}\,\{\lambda v_t v\} \le \lambda |v_t|^2 - m(t)\lambda^3 |v|^2 + n(t)\lambda^2 |\psi||v| + n(t)\lambda|\mu||\psi||v|. \tag{3.4}$$

We will suppose that $q(t) \geq q_0$ (otherwise we take $-\psi v_t$ instead of ψv_t). Multiplying (3.2) by v_t we obtain

$$\frac{d}{dt}\mathrm{Re}\{\psi v_t\} = -r(t)(\lambda+\alpha)\mathrm{Re}\{\psi v_t\} + q(t)\lambda|v_t|^2 + \mu q(t)|v_t|^2 - m(t)\lambda^2\mathrm{Re}\{v\psi\} - n(t)\lambda|\psi|^2 - n(t)\mu|\psi|^2. \tag{3.5}$$

Inequalities (3.4) and (3.5) imply

$$\frac{d}{dt}\mathrm{Re}\{\lambda v_t v\} \leq \lambda|v_t|^2 - \frac{m_0}{2}\lambda^3|v|^2 + \frac{n^2(t)}{2m_0}\lambda|\psi|^2 + n(t)\lambda|\mu||\psi||v|, \tag{3.6}$$

$$\frac{d}{dt}\mathrm{Re}\{-\psi v_t\} \leq -\frac{q_0}{2}\lambda|v_t|^2 + \frac{r^2(t)\lambda}{2q_0}|\psi|^2 + \{r(t)|\alpha| - q(t)\mu\}|v_t|^2 + \frac{q_0 m_0}{16}\lambda^3|v|^2 + \frac{4m^2(t)}{q_0 m_0}\lambda|\psi|^2 + n(t)\lambda|\psi|^2 + (n(t)|\mu| + r(t)|\alpha|)|\psi|^2, \tag{3.7}$$

respectively. From (3.6), (3.7) we conclude

$$\frac{d}{dt}\mathrm{Re}\left\{\frac{q_0}{4}\lambda v_t v - \psi v_t\right\} \leq -\frac{q_0}{4}\lambda|v_t|^2 - \frac{q_0 m_0 \lambda^3}{16}|v|^2 + c\lambda|\psi||v| + c|v_t|^2 + c\lambda|\psi|^2 + c|\psi|^2 \tag{3.8}$$

with c being a constant depending essentially on T, possibly varying from formula to formula. Combining (3.3) and (3.8) we obtain

$$\frac{d}{dt}\mathcal{K}(t,\lambda) \leq -c\lambda\left\{|v_t(t,\lambda)|^2 + \lambda^2|v(t,\lambda)|^2 + |\psi(t,\lambda)|^2\right\} + c\left\{\lambda^2|v(t,\lambda)|^2 + |\psi(t,\lambda)|^2 + |v_t(t,\lambda)|^2\right\}, \tag{3.9}$$

where

$$\mathcal{K}(t,\lambda) := \mathcal{E}(t,\lambda) + \varepsilon\frac{q_0}{4}\lambda\mathrm{Re}\{v_t v\} - \varepsilon\mathrm{Re}\{\psi v_t\}.$$

Taking ε small enough, we get

$$\frac{1}{2}\mathcal{E}(t,\lambda) \leq \mathcal{K}(t,\lambda) \leq 2\mathcal{E}(t,\lambda), \tag{3.10}$$

hence

$$\frac{d}{dt}\mathcal{K}(t,\lambda) \leq -c_1\lambda\mathcal{K}(t,\lambda) + c_2\mathcal{K}(t,\lambda)$$

with positive constants c_1, c_2. We will consider two cases. First, we consider the case in which $\lambda \geq 2c_2/c_1 =: c_3$; then we consider the case in which $\lambda \leq c_3$. For $\lambda \geq c_3$ we get

$$\frac{d}{dt}\mathcal{K}(t,\lambda) \leq -\frac{c_1}{2}\lambda\mathcal{K}(t,\lambda);$$

thus

$$\mathcal{E}(t,\lambda) \leq c\mathcal{E}(0,\lambda)e^{-\frac{c_1}{2}\lambda t}. \tag{3.11}$$

Multiplying by λ^m and integrating for $\lambda \geq c_3$, we get

$$\int_{\lambda \geq c_3} \lambda^m \mathcal{E}(\lambda, t)\, d\mu(\lambda) \leq \int_{\lambda \geq c_3} \lambda^m \mathcal{E}(\lambda, 0) e^{-\frac{c_1}{2}\lambda t}\, d\mu(\lambda). \tag{3.12}$$

On the other hand, if $\lambda < c_3$ we get

$$\frac{d}{dt}\mathcal{K}(t,\lambda) \leq c_2 \mathcal{K}(t,\lambda) \Rightarrow \mathcal{K}(\lambda, t) \leq e^{c_2 t}\mathcal{K}(\lambda, 0) \quad \forall \lambda \in [0, c_3].$$

Using (3.10) we get

$$\mathcal{E}(t,\lambda) \leq c\mathcal{E}(0,\lambda)e^{c_2 t}. \tag{3.13}$$

Multiplying by λ^m and integrating over $0 \leq \lambda \leq c_3$, we obtain

$$\int_{\lambda \leq c_3} \lambda^m \mathcal{E}(\lambda, t)\, d\mu(\lambda) \leq c c_3^m \int_{\lambda \leq c_3} \mathcal{E}(\lambda, 0) e^{c_2 t}\, d\mu(\lambda). \tag{3.14}$$

Finally, from (3.12) and (3.14) we conclude that for $t > 0$,

$$\int_0^\infty \lambda^m \mathcal{E}(\lambda, t)\, d\mu(\lambda) \leq c(t,m) \int_0^\infty \mathcal{E}(\lambda, 0)\, d\mu(\lambda).$$

Using the diagonalization theorem (cf. [2]), we get

$$\forall t > 0: \ ||A^m u(t)|| + ||A^m \theta(t)|| \leq c(t,m)\left\{||u_1|| + ||Au_0|| + ||A\theta_0||\right\}. \tag{3.15}$$ □

Remark 3.2. The constant $c(m,t)$, given in inequality (3.15), is such that $c(m,t) \to \infty$ as $t \to 0$.

Remark 3.3. If M, N, R, and Q are C^{k-1}-functions, then the solution (u,θ) of (1.1),(1.2) satisfies

$$(u,\theta) \in C^k(]0,T]; \cap_{j \in \mathbb{N}} D(A^j)).$$

Remark 3.4. The smoothness effect property does not depend on the largeness of the initial data, because the method we used can be applied for local or global solutions.

THEOREM 3.5. *Let $A \geq \nu > 0$, $\alpha \geq 0$, $\mu \geq 0$, and let $(u,\theta) \in \cap_{j=0}^2 C^j([0,\infty], D(A^{2-j})) \times \cap_{j=0}^1 C^j([0,\infty], D(A^{1-j}))$ be a solution to* (1.12), (1.13), (1.3), (1.4). *Then (u,θ) decays to zero exponentially, i.e.,*

$$E(t) \leq Me^{-dt}E(0)$$

for some positive constants M, d, where

$$E(t) := \frac{1}{2}\left\{||u_t(t)||^2 + m(t)||Au(t)||^2 + \frac{n(t)}{q(t)}||\theta(t)||^2\right\}.$$

Proof. With the same technique as in the proof of Theorem 3.1—the energy method—we conclude from (3.11) that there are $M > 0$ and $c_1 > 0$ for which we have

$$E(t) = \int_\nu^\infty \mathcal{E}(t,\lambda)\, d\mu(\lambda) \leq M \int_\nu^\infty e^{-c_1 \lambda t}\mathcal{E}(0,\lambda)\, d\mu(\lambda) \leq Me^{-c_1 \nu t}E(0).$$ □

We observe that c_1 depends on the C^1-norm of m, n, and q.

When the operator A is not coercive, that is, $A \geq 0$ only, the exponential decay is not expected. In the following theorem we will study this case when $\Omega = \mathbb{R}^n$ and $\Omega = \mathbb{R}^n \setminus B$, where B is a bounded closed set.

THEOREM 3.6. *Let $\Omega = \mathbb{R}^n$ or let $n \geq 3$ and $\Omega = \mathbb{R}^n \setminus B$, where $B \neq \emptyset$ is a bounded closed set with smooth boundary, and let $\mathbb{R}^n \setminus \Omega$ be star shaped. Then, for the solution (u, θ) of (1.5)–(1.7), (1.3), we have that*

$$||(u_t, \Delta u, \theta)(t)||_{L^\infty(\Omega)\{L^2(\Omega)\}} \leq c t^{-\frac{n}{2}\{-\frac{n}{4}\}} ||(u_1, \Delta u_0, \theta_0)(t)||_{L^1(\Omega)}$$

with a positive constant c neither depending on t nor on the initial data.

Proof. First let $\Omega = \mathbb{R}^n$. Denoting by $\hat{u}(t, \xi)$ and $\hat{\theta}(t, \xi)$ the Fourier transform of u and θ, respectively, we obtain

$$\hat{u}_{tt}(t, \xi) + |\xi|^4 \hat{u}(t, \xi) - \beta |\xi|^2 \hat{\theta}(t, \xi) = 0, \tag{3.16}$$

$$\hat{\theta}_t(t, \xi) + |\xi|^2 \hat{\theta}(t, \xi) + \beta |\xi|^2 \hat{u}_t(t, \xi) = 0. \tag{3.17}$$

Combining (3.16), (3.17) with (3.1), (3.2), and defining

$$\hat{\mathcal{E}}(t, \xi) := \frac{1}{2} \left\{ |\hat{u}_t(t, \xi)|^2 + |\xi|^4 |\hat{u}(t, \xi)|^2 + |\hat{\theta}(t, \xi)|^2 \right\},$$

we obtain, by the same multiplicative technique as in the proof of Theorem 3.1,

$$\exists M > 0 \; \exists d > 0 \; \forall t \geq 0 \; \forall \xi \in \mathbb{R}^n : \hat{\mathcal{E}}(t, \xi) \leq M e^{-d|\xi|^2 t} \hat{\mathcal{E}}(0, \xi),$$

which implies for

$$\mathcal{E}(t) := \frac{1}{2} \left\{ |u(t)|^2 + |\Delta u(t)|^2 + |\theta(t)|^2 \right\},$$

$$\begin{aligned} \mathcal{E}(t, \lambda) &= \int_{\mathbb{R}^n} \hat{\mathcal{E}}(t, \xi) \, d\xi \leq M \int_{\mathbb{R}^n} e^{-d|\xi|^2 t} \hat{\mathcal{E}}(0, \xi) \, d\xi \\ &\leq c t^{-\frac{n}{2}} ||\hat{\mathcal{E}}(0, \xi)||_{L^\infty_\xi} \leq c t^{-\frac{n}{2}} ||(u_1, \Delta u_0, \theta_0)||_{L^1(\Omega)}. \end{aligned} \tag{3.18}$$

Moreover,

$$\begin{aligned} |u_t(x, t)| &\leq \left| \left\{ \frac{1}{\sqrt{2\pi}} \right\}^n \int_{\mathbb{R}^n} e^{ix\xi} \hat{u}_t(t, \xi) \, d\xi \right| \\ &\leq c \int_{\mathbb{R}^n} e^{-\frac{d}{2}|\xi|^2 t} \sqrt{\hat{\mathcal{E}}(0, \xi)} \, d\xi \leq c t^{-\frac{n}{2}} ||(u_1, \Delta u_0, \theta_0)||_{L^1(\Omega)}. \end{aligned} \tag{3.19}$$

(3.18) and (3.19) prove Theorem 3.6 for $\Omega = \mathbb{R}^n$. Now, let $\Omega \neq \mathbb{R}^n$ be an exterior domain $n \geq 3$. There exists a generalized Fourier transform $\mathcal{F} : L^2(\Omega) \to L^2(\mathbb{R}^n)$ such that

$$\mathcal{F}(\varphi(A)w)(\xi) = \varphi(|\xi|^2)(\mathcal{F}w)(\xi), \tag{3.20}$$

where A is the Laplace operator defined on $H_0^1(\Omega) \cap H^2(\Omega)$ and $\varphi(A)$ is assumed to be defined via the spectral theorem. $\mathcal{F}$ is represented by

$$(\mathcal{F}w)(\xi) = \int_\Omega \hat{\psi}(x, \xi) w(x) \equiv \hat{w}(t, \xi),$$

$$(\mathcal{F}^{-1}\hat{w})(x) = \int_{\mathbb{R}^n} \psi(x,\xi)\hat{w}(\xi)\,d\xi$$

with a kernel $\psi(x,\xi)$; see [14], [17]. In [17] it is proved, based on results from [9], that

$$\exists m \in N\ \exists c > 0\ \exists x \in \Omega\ \forall \xi \in \mathbb{R}^n \setminus \{0\}: \quad |\psi(x,\xi)| \le c(1+|\xi|)^m \tag{3.21}$$

holds, provided $\mathbb{R}^n \setminus \Omega$ is star shaped. Using (3.20), we obtain the analogue of (3.16), (3.17). Essentially repeating the calculation following (3.17), we obtain (3.18) again and, using (3.19), we get

$$|u_t(t,x)| \le c\int_{\mathbb{R}^n} e^{-\frac{d}{2}|\xi|^2 t}\sqrt{\mathcal{E}(0,\lambda)}(1+|\xi|)^m \le ct^{-\frac{n}{2}}||(u_1,\Delta u_0,\theta_0)||_{L^1(\Omega)}. \qquad \square$$

In one space dimension we can use the Fourier-sine transform [7], for example, if $\Omega =]0,\infty[$, to obtain the corresponding result. For $n = 2$ the known estimate for $\psi(x,\xi)$ has a factor $\log|\xi|$ as $|\xi| \to 0$, which leads to a decay like $c_\varepsilon t^{-n/4+\varepsilon}$ and $c_\varepsilon t^{-n/2+\varepsilon}$, respectively (instead of $ct^{-1/2}$ and ct^{-1} as expected).

4. Propagation of singularities.

THEOREM 4.1. *Let $A \ge 0$ be self-adjoint in a separable Hilbert space $\mathcal{H}$ and let v be a solution to* (1.23)–(1.25). *Then we have for $v_0 = v_2 = 0$, that*

$$\forall s \ge 0 : v_1 \notin D(A^{s+2}) \Rightarrow \forall t \ge 0 : \ \lambda \mapsto \lambda^{s+2}\left(\mathcal{U}v_t(t,\lambda), \lambda^{\frac{1}{2}}\mathcal{U}v(t,\lambda)\right) \notin \mathcal{H}\times\mathcal{H}.$$

Remark 4.2. In terms of the example from one-dimensional thermoelasticity, the nonsmoothing of the "hyperbolic"energy $\|u_t(t)\|^2 + \|u_x(t)\|^2$ is proved. The formulation in terms of v_1 is made for simplicity of the exposition; similar results could be obtained in terms of v_0 and v_2.

Proof of Theorem 4.1. Using the spectral theorem, we conclude from (1.23), (1.24) that $w(t,\lambda) := \mathcal{U}v(t,\lambda)$ satisfies

$$w_{ttt} + \kappa\lambda w_{tt} + (\gamma^2+\tau)\lambda w_t + \kappa\tau\lambda^2 w = 0, \tag{4.1}$$

$$w(t=0) = w_0 := \mathcal{U}v_0, \quad w_t(t=0) = w_1 := \mathcal{U}v_1, \quad w_{tt}(t=0) = w_2 := \mathcal{U}v_2. \tag{4.2}$$

The solution w of (4.1), (4.2) is given by

$$w(t,\lambda) = \sum_{j=1}^{3} b_j(\lambda)e^{-\beta_j(\lambda)t},$$

where $\beta_j(\lambda)$, $j = 1,2,3$, are the roots of the characteristic equation

$$-\beta^3 + \kappa\lambda\beta^2 - (\gamma^2+\tau)\lambda\beta + \kappa\tau\lambda^2 = 0$$

and

$$b_j(\lambda) := \sum_{k=0}^{2} b_j^k(\lambda)w_k(\lambda)$$

with

$$b_j^0 := \frac{\prod_{l\neq j}\beta_l}{\prod_{l\neq j}(\beta_j-\beta_l)}, \quad b_j^1 := \frac{\sum_{l\neq j}\beta_l}{\prod_{l\neq j}(\beta_j-\beta_l)}, \quad b_j^2 := \frac{1}{\prod_{l\neq j}(\beta_j-\beta_l)}.$$

Since $w_0 = w_2 = 0$, we obtain

$$w(t,\lambda) = \sum_{j=1}^{3} b_j^1 e^{-\beta_j(\lambda)t} w_1(\lambda) \equiv \sum_{j=1}^{3} f^j(t,\lambda).$$

The asymptotic behavior of $\beta_j(\lambda)$ is known (see [19], [15]) and given as follows.

LEMMA 4.3. *As* $\lambda \to 0$,

$$\beta_1(\lambda) = \frac{\kappa\tau}{\tau+\gamma^2}\lambda + O(\lambda^{\frac{3}{2}}), \quad \beta_{\frac{2}{3}}(\lambda) = \frac{\kappa\gamma^2}{2(\tau+\gamma^2)}\lambda \pm i\sqrt{\tau+\gamma^2}\sqrt{\lambda} + O(\lambda^{\frac{3}{2}});$$

as $\lambda \to \infty$,

$$\beta_1(\lambda) = \kappa\lambda - \frac{\gamma^2}{\kappa} - \frac{\alpha_1}{\kappa^3}\lambda^{-1} + O(\lambda^{-\frac{3}{2}}),$$

$$\beta_{\frac{2}{3}}(\lambda) = \frac{\gamma^2}{2\kappa} + \frac{\alpha_2}{2\kappa^3}\lambda^{-1} + O(\lambda^{-2}) \mp i\left\{\sqrt{\tau\lambda} + \frac{\alpha_3}{\kappa_2}\lambda^{-\frac{1}{2}} + O(\lambda^{-\frac{3}{2}})\right\},$$

where $\alpha_j = \alpha_j(\gamma,\tau)$ *are constants* $(j = 1,2)$.

Except for at most two values of $\lambda > 0$, *we get*

$$\beta_j(\lambda) \neq \beta_k(\lambda), \quad j \neq k.$$

For any value of $\lambda \neq 0$, $\mathrm{Re}\beta_j(\lambda) > 0$, $j = 1,2,3$.

There are positive constants r_1 *and* C_j, $j = 1,2,3$, *such that*

$$\lambda \leq r_1^2 \Rightarrow C_1\lambda \leq \mathrm{Re}\beta_j(\lambda) \leq C_2\lambda,$$

$$\lambda \geq r_1^2 \Rightarrow \mathrm{Re}\beta_j(\lambda) \geq C_3 \ (j = 1,2,3).$$

This implies the following asymptotic behavior for $b_j^1(\lambda)$ $j = 1,2,3$.

LEMMA 4.4. *As* $\lambda \to 0$,

$$b_1^1(\lambda) = \kappa\gamma^2 + O(\sqrt{\lambda}), \quad b_{\frac{2}{3}}^1(\lambda) = \frac{\pm i}{2\sqrt{\lambda(\tau+\gamma^2)}} + O(1);$$

as $\lambda \to \infty$,

$$b_1^1(\lambda) = O(\lambda^{-2}), \quad b_{\frac{2}{3}}^1(\lambda) = \frac{\mp i}{2\sqrt{\lambda}} + O\left(\frac{1}{\lambda}\right).$$

Observe that the leading term for $b_{2/3}^1(\lambda)$ as $\lambda \to 0$ is like $\lambda^{-1/2}$ but, still, $b_1^1 e^{-\beta_2 t} + b_3^1 e^{-\beta_3 t} = O(1)$ as $\lambda \to 0$; hence the interesting part is $\lambda \to \infty$. Now let $t > 0$. It is easy to see that for any $m \in \mathbb{N}$,

$$\infty > \int_0^\infty \lambda^{2m}\left\{|f_t^1(t,\lambda)|^2 + \lambda|f^1(t,\lambda)|^2\right\} d\mu(\lambda).$$

Hence the f^1-part is arbitrarily smooth. We will now prove that the remaining part of w it is not smoother than w_1. In fact, let us suppose the contrary, so we have

$$\infty > \int_{r_1}^{\infty} \lambda^{2s+4} \left\{ |f_t^2(t,\lambda) + f_t^3(t,\lambda)|^2 + \lambda |f^2(t,\lambda) + f^3(t,\lambda)|^2 \right\} d\mu(\lambda).$$

We obtain for $r_1 > 0$ sufficiently large, depending on t later on using Lemmas 4.3 and 4.4,

$$\begin{aligned}
\infty &> \int_{r_1}^{\infty} \lambda^{2s+4} \left\{ \left| \frac{i}{2\sqrt{\lambda}} (\beta_2 e^{-\beta_2 t} - \beta_3 e^{-\beta_3 t}) \right|^2 + O\left(\frac{1}{\lambda^2}\right) \right. \\
&\qquad \left. + \lambda \left| \frac{i}{2\sqrt{\lambda}} (e^{-\beta_3 t} - e^{-\beta_2 t}) \right|^2 + O\left(\frac{1}{\lambda}\right) \right\} |w_1(\lambda)|^2 \, d\mu(\lambda) \\
&= \int_{r_1}^{\infty} \lambda^{2s+4} \frac{e^{-2at}}{4} \left\{ \left| \sqrt{\tau} \cos(bt) + O\left(\frac{1}{\sqrt{\lambda}}\right) \right|^2 + |\sin(bt)|^2 + O\left(\frac{1}{\lambda}\right) \right\} |w_1(\lambda)|^2 \, d\mu(\lambda),
\end{aligned}$$

where $a := \operatorname{Re}\beta$, $b := \operatorname{Im}\beta$. Thus we obtain, for $r_1 = r_1(t)$ sufficiently large,

$$\infty > \frac{\min\{\tau, 1\}}{4} e^{-\frac{\gamma^2 t}{2\kappa}} \int_{r_1(t)}^{\infty} \lambda^{2s+4} |w_1(\lambda)|^2 \, d\mu(\lambda),$$

which is a contradiction because $v_1 \notin D(A^{s+2})$. □

Acknowledgment. The authors thank Y. Shibata for discussions concerning the results of this paper.

REFERENCES

[1] R. A. ADAMS, *Sobolev spaces*, Academic Press, New York, 1975.

[2] D. HUET, *Décomposition spectrale et opérateurs*, Presses Universitaires de France, Nancy, 1976.

[3] S. JIANG, *Global existence of smooth solutions in one-dimensional nonlinear thermoelasticity*, Proc. Roy. Soc. Edinburgh, Sect. A, 115 (1990), pp. 257–274.

[4] T. KATO, *Abstract differential equations and nonlinear mixed problems*, in Lezioni Fermiane, Accad. Naz. D. Linc. Scuola Norm. Sup. Pisa, 1985.

[5] S. KAWASHIMA AND M. OKADA, *Smooth global solutions for the one-dimensional equations in magnetohydrodynamics*, Proc. Jap. Acad., Ser. A, 53 (1982), pp. 384–387.

[6] J. U. KIM, *On the energy decay of a linear thermoelastic bar and plate*, SIAM J. Math. Anal., 23 (1992), pp. 889–899.

[7] R. LEIS, *Initial boundary value problems in mathematical physics*, Teubner-Verlag, Stuttgart, John Wiley, Chichester, 1986.

[8] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, vol. 1. Dunod, Paris (1968).

[9] C. S. MORAWETZ AND D. LUDWIG, *An inequality for the reduced wave operator and the justification of geometrical optics*, Comm. Pure Appl. Math., 21 (1968), pp. 187–203.

[10] J. E. MUÑOZ RIVERA, *Energy decay rates in linear thermoelasticity*, Funkcial. Ekvac., 35 (1992), pp. 19–30.

[11] ———, *Decomposition of the displacement vector field and decay rates in linear thermoelasticity*, SIAM J. Math. Anal., 24 (1993), pp. 390–406.

[12] R. RACKE, *On the Cauchy problem in nonlinear 3-d-thermoelasticity*, Math. Z., 203 (1990), pp. 649–682.

[13] ———, *L^p-L^q-estimates for solutions to the equations of linear thermoelasticity in exterior domains*, Asymptotic Anal., 3 (1990), pp. 105–132.

[14] ———, *Decay rates for solutions of damped systems and generalized Fourier transforms*, J. Reine Angew. Math., 412 (1990), pp. 1–19.

[15] ———, *Lectures on nonlinear evolution equations. Initial value problems*, in Aspects of Mathematics E19, Friedr. Vieweg Sohn, Braunschweig and Wiesbaden, 1992.

[16] R. RACKE, *Mathematical aspects in thermoelasticity*, in Sonderforschungsbereich SFB 256 Vorlesungsreihe 25, Universität Bonn, 1992.

[17] R. RACKE AND S. ZHENG, *Global existence of solutions to a fully nonlinear fourth-order parabolic equation in exterior domains*, Nonlinear Anal., 17 (1992), pp. 1027–1038.

[18] M. SLEMROD, *Global existence, uniqueness, and asymptotic stability of classical smooth solutions in one-dimensional non-linear thermoelasticity*, Arch. Rational Mech. Anal., 76 (1981), pp. 97–133.

[19] S. ZHENG AND W. SHEN, *Global solutions to the Cauchy problem of quasilinear hyperbolic parabolic coupled systems*, Sci. Sinica Ser. A, 30 (1987), pp. 1133–1149.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1564–1587, November 1995

011

PERTURBED SCALE-INVARIANT INITIAL VALUE PROBLEMS IN ONE-DIMENSIONAL DYNAMIC ELASTOPLASTICITY*

MICHAEL K. GORDON†

Abstract. The author considers an initial value problem for equations describing the longitudinal motion of an elastoplastic rod. Conditions on the stress σ determine whether the deformation of the rod is *plastic* or *elastic,* both of which are described by wave equations with different wave speeds. Also, plastic deformation is quasi-linear while elastic deformation is assumed to be linear. The initial conditions are continuous, piecewise C^1, and have a jump in the first derivative only at the origin. This is a generalization of the *scale-invariant* problem solved by D. Schaeffer and M. Shearer, in which plastic deformation is assumed to be linear and the initial conditions are piecewise linear.

The analysis is divided into cases according to the structure of the corresponding scale-invariant problem; the most interesting case reduces to a free boundary problem for the plastic equations on a wedge with two free boundaries.

Key words. scale-invariant problem, plastic and elastic regions, free boundary problem, local existence

AMS subject classifications. 35L60, 35R35

Introduction. Elastoplasticity is an important subject in the study of nonlinear stress-strain responses in solid mechanics. In this work, we consider the following one-dimensional elastoplasticity model:

$$\begin{aligned} \partial_t v &= \partial_x \sigma, \\ \partial_t \sigma + k(\gamma)\partial_t \gamma &= \partial_x v, \\ \partial_t \gamma &= \begin{cases} 0 & \text{if } \sigma < \gamma, \\ (\partial_t \sigma)_+ & \text{if } \sigma = \gamma, \end{cases} \end{aligned} \tag{0.1}$$

where v is velocity, σ is stress, γ is yield stress (the stress at which the material deforms plastically), and k is a smooth (smooth will mean C^1 throughout), positive function. Notice that $\sigma \leq \gamma$ for any solution of (0.1). These equations are a simplified version of equations which describe longitudinal motion of an elastoplastic rod with hardening (cf. Lee [4], Clifton and Bodner [2], and Antman and Szymczak [1]).

We consider (0.1) with the following piecewise smooth initial data.

$$\begin{aligned} v(x,0) &= \begin{cases} a_L x + f(x) & \text{for } x \leq 0, \\ a_R x + f(x) & \text{for } x \geq 0, \end{cases} \\ \sigma(x,0) &= \begin{cases} b_L x + g(x) & \text{for } x \leq 0, \\ b_R x + g(x) & \text{for } x \geq 0, \end{cases} \\ \gamma(x,0) &= \begin{cases} c_L x + h(x) & \text{for } x \leq 0, \\ c_R x + h(x) & \text{for } x \geq 0, \end{cases} \end{aligned} \tag{0.2}$$

* Received by the editors September 1, 1993; accepted for publication (in revised form) March 3, 1994. This research was supported in part by National Science Foundation grants DMS 8804592 and DMS 9201034, which include funds from the Air Force Office of Scientific Research.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

where f, g, and h are smooth and they and their first derivatives vanish at the origin. We also require that $\sigma(x,0) \leq \gamma(x,0)$. In particular, this means that $c_L \leq b_L$ and $b_R \leq c_R$.

Equation (0.1) may also be written in the following form:

$$
\begin{aligned}
&\text{(a)} \quad \partial_t v = \partial_x \sigma, \\
&\qquad \partial_t \sigma = c(\sigma)^2 \partial_x v, \quad \text{if } \sigma = \gamma \quad \text{and} \quad \partial_t \sigma \geq 0, \\
&\qquad \partial_t \gamma = \partial_t \sigma; \\
&\text{(b)} \quad \partial_t v = \partial_x \sigma, \\
&\qquad \partial_t \sigma = \partial_x v, \quad \text{if } \sigma < \gamma \quad \text{or} \quad \partial_t \sigma \leq 0, \\
&\qquad \partial_t \gamma = 0,
\end{aligned}
\tag{0.3}
$$

where $c(\sigma) = (1 + k(\sigma))^{-1/2}$. We say that the solution is *plastic* wherever (0.3(a)) holds and *elastic* wherever (0.3(b)) holds. Notice that both cases of (0.3) are satisfied if $\sigma = \gamma$ and $\partial_t \sigma = 0$.

In [5], Schaeffer and Shearer consider (0.1), (0.2) with constant k and $f, g, h \equiv 0$, and they construct a unique, continuous, piecewise linear similarity solution, consisting of wedges emanating from the origin on which the solution is linear and either plastic or elastic. They refer to this as the *scale-invariant* problem since it is invariant under the scaling

$$\tilde{U}(x,t) = \eta^{-1} U(\eta x, \eta t).$$

Recall that the Riemann problem has piecewise constant initial data with a jump only at the origin, and admits solutions which are invariant under the scaling $\tilde{U}(x,t) = U(\eta x, \eta t)$. The basis of the construction in [5] is to study the jumps in the first derivatives between wedges in the same way that shocks and rarefaction waves are studied in solving the Riemann problem. In [6], The Riemann problem is generalized by taking piecewise smooth initial data with a discontinuity at the origin; (0.1), (0.2) can be regarded as an analogous generalization of the scale-invariant problem.

In §1 we show that, under certain genericity assumptions, the solution to (0.1), (0.2) must be of the same form as the corresponding scale-invariant solution. In §§2, 3, and 4, we solve the initial value problem (0.1), (0.2) locally (near $x = 0$ for small time) by dividing the solution into cases according to the form of the corresponding scale-invariant solution. In each case, the problem is reduced to a boundary value problem in the xt-plane for either the plastic equations or the elastic equations. In §2, we derive the most interesting of these: a free boundary problem for the plastic equations on a wedge with two free boundaries, which we solve in §4. The purpose of §3 is to solve certain fixed boundary problems which are needed to solve the free boundary problem.

1. The form of the solutions. We begin by defining a class of piecewise smooth functions in which we expect to find solutions.

DEFINITION 1.1. *Let $\mathcal{U}$ be the set of continuous functions*

$$U(x,t) = \begin{pmatrix} v(x,t) \\ \sigma(x,t) \\ \gamma(x,t) \end{pmatrix}, \qquad \sigma \leq \gamma,$$

where U is defined on some $t \geq 0$ neighborhood of the origin and either U is smooth or its domain can be partitioned into sectors in the following way. There are smooth functions $\{\eta_i(t)\}_{i=1}^m$ such that

(i) $\eta_i(0) = 0$,
(ii) $\eta_i(t) < \eta_{i+1}(t)$, $t > 0$,
(iii) U *is smooth on*

$$T_i = \{(x,t): t \geq 0, \eta_i(t) \leq x \leq \eta_{i+1}(t)\}, \quad 1 \leq i \leq m-1,$$
$$T_0 = \{(x,t): t \geq 0, x \leq \eta_1(t)\}, \quad \textit{and}$$
$$T_m = \{(x,t): t \geq 0, x \geq \eta_m(t)\},$$

(iv) U *is not smooth across any* $\eta_i(t)$.

Define

$$U_i(x,t) = \begin{pmatrix} v_i \\ \sigma_i \\ \gamma_i \end{pmatrix} = U|_{T_i}.$$

DEFINITION 1.2. *Let* $U \in \mathcal{U}$ *satisfy* (0.2). *Define the wedges*

$$\bar{T}_i = \{(x,t): t \geq 0, \eta_i'(0)t \leq x \leq \eta_{i+1}'(0)t\}, \qquad 1 \leq i \leq m-1,$$
$$\bar{T}_0 = \{(x,t): t \geq 0, x \leq \eta_1'(0)t\},$$
$$\bar{T}_m = \{(x,t): t \geq 0, x \geq \eta_m'(0)t\},$$

and define $\bar{U}(x,t)$ *by*

$$\bar{U}|_{\bar{T}_i} = \bar{U}_i = \begin{pmatrix} \bar{v}_i \\ \bar{\sigma}_i \\ \bar{\gamma}_i \end{pmatrix} = A_i x + B_i t,$$

where $(A_i, B_i) = \lim_{(x,t)\to(0,0)} \nabla U_i$. *Notice that* $\bar{U}$ *is simply the linearization of* U *at the origin and that the boundaries of the regions* $\bar{T}_i$ *are tangent to the boundaries of* T_i *at the origin.*

DEFINITION 1.3. *For* $t > 0, 1 \leq i \leq m$, *define*

$$U_{i,x}^{\pm}(t) = \lim_{x \to \eta_i(t)^{\pm}} \partial_x U(x,t),$$
$$U_{i,t}^{\pm}(t) = \lim_{x \to \eta_i(t)^{\pm}} \partial_t U(x,t).$$

From the continuity of U, we have

$$(U_{i,x}^{+}(t) - U_{i,x}^{-}(t))\eta_i'(t) + (U_{i,t}^{+}(t) - U_{i,t}^{-}(t)) = 0. \tag{1.1}$$

Letting $t \to 0^+$, we have

$$(\partial_x \bar{U}_i - \partial_x \bar{U}_{i-1})\eta_i'(0) + (\partial_t \bar{U}_i - \partial_t \bar{U}_{i-1}) = 0, \tag{1.2}$$

which implies that $\bar{U}_i$ and $\bar{U}_{i-1}$ agree on $x = \eta_i'(0)t$.

If U is a solution of (0.1), (0.2) then $\bar{U}$ solves the corresponding scale-invariant problem. More precisely, we have the following lemma.

LEMMA 1.1. *Let* $U \in \mathcal{U}$ *satisfy* (0.1), (0.2). *Then* $\bar{U}$ *is the unique, continuous, piecewise linear solution (given in* [5]*) to* (0.1), (0.2) *with* $k \equiv k(0)$ *and* $f, g, h \equiv 0$.

Proof. It is clear from (1.2) and Definition 1.2 that $\bar{U}$ is continuous and satisfies (0.2) with $f, g, h \equiv 0$. Hence, we just need to show that $\bar{U}_i$ satisfies (0.1) with $k \equiv k(0)$

on any $\bar{T}_i$ with nonempty interior, i.e., $\eta_i'(0) \neq \eta_{i+1}'(0)$. It is easy to see from (0.3) that $\bar{U}_i$ must satisfy one of the following:

$$
\begin{aligned}
&\text{(a)} \quad \partial_t \bar{v}_i = \partial_x \bar{\sigma}_i, \\
&\qquad\quad \partial_t \bar{\sigma}_i = c_0^2 \partial_x \bar{v}_i, \\
&\qquad\quad \partial_t \bar{\gamma}_i = \partial_t \bar{\sigma}_i \geq 0; \\
&\text{(b)} \quad \partial_t \bar{v}_i = \partial_x \bar{\sigma}_i, \\
&\qquad\quad \partial_t \bar{\sigma}_i = \partial_x \bar{v}_i, \\
&\qquad\quad \partial_t \bar{\gamma}_i = 0,
\end{aligned}
\tag{1.3}
$$

where $c_0 = (1 + k(0))^{-1/2}$.

Suppose $\partial_x \bar{v}_i \neq 0$. Then only one of (a) or (b) holds. If (a), then U_i is plastic locally (near the origin), so $\bar{\sigma}_i = \bar{\gamma}_i$ and $\partial_t \bar{\sigma}_i \geq 0$. If (b), then $\partial_t \bar{\sigma}_i = \partial_x \bar{v}_i \neq \partial_t \bar{\gamma}_i$. But $\bar{\sigma}_i \leq \bar{\gamma}_i$ since $\sigma_i \leq \gamma_i$, so $\bar{\sigma}_i < \bar{\gamma}_i$ on the interior of $\bar{T}_i$ because $\partial_t \bar{\sigma}_i \neq \partial_t \bar{\gamma}_i$. Hence $\bar{U}_i$ satisfies (0.1) with $k \equiv k(0)$ in either case.

If $\partial_x \bar{v}_i = 0$ then both (a) and (b) hold, so $\bar{U}_i$ satisfies (0.1) with $k \equiv k(0)$ regardless of whether $\bar{\sigma}_i \equiv \bar{\gamma}_i$ or $\bar{\sigma}_i < \bar{\gamma}_i$ on the interior of $\bar{T}_i$. This completes the proof. □

As mentioned in the introduction, we would like to be able to say that a solution U of (0.1), (0.2) is locally of the *same form* as $\bar{U}$. By this we mean that U_i is of the same type (plastic or elastic) as $\bar{U}_i$ in a neighborhood of the origin and that no $\bar{U}_i$ has empty interior. In order for this to be true, we must avoid "borderline" cases of the scale-invariant problem. Also, we consider only one of any two cases which give solutions which are reflections across the t-axis of one another. This leads to the assumption that one of the following holds.

$$
\begin{aligned}
&\text{(i)} \quad s_L, s_R < 1, \qquad b_L - a_L > b_R + a_R, \\
&\text{(ii)} \quad s_L, s_R < 1, \qquad b_L - a_L < b_R + a_R, \\
&\text{(iii)} \quad s_L, s_R > 1, \\
&\qquad\quad b_L - a_L(1 + c_0 s_L)/(c_0 + s_L) > b_R + a_R(1 + c_0 s_R)/(c_0 + s_R), \\
&\text{(iv)} \quad s_L, s_R > 1, \\
&\qquad\quad b_L - a_L(1 + c_0 s_L)/(c_0 + s_L) < b_R + a_R(1 + c_0 s_R)/(c_0 + s_R), \\
&\text{(v)} \quad s_R < 1 < s_L, \\
&\qquad\quad b_L - a_L(1 + c_0 s_L)/(c_0 + s_L) > b_R + a_R, \\
&\text{(vi)} \quad s_R < 1 < s_L, \\
&\qquad\quad b_L - a_L(1 + c_0 s_L)/(c_0 + s_L) < b_R + a_R,
\end{aligned}
\tag{1.4}
$$

where $s_L = a_L/(b_L - c_L)$, $s_R = a_R/(c_R - b_R)$, and $c_0 = (1 + k(0))^{-1/2}$.

Replacing inequality with equality anywhere in (1.4) would represent a borderline case. In borderline cases, the solution to the scale-invariant problem may be both plastic and elastic on some region or yielding on a boundary between two elastic regions. In such a situation, the local form of the solution to the perturbed scale-invariant problem (0.1), (0.2) would depend on f, g, and h, a complication which we avoid here. We shall refer to initial data (0.2) satisfying one case of (1.4) as *nondegenerate*.

It is shown in [5] that if the initial data (0.2) is nondegenerate then $\bar{U}$ is of one of the forms shown in Fig. 1.1. $\bar{U}$ is plastic on P_i and elastic on E_i. The arrows in the figure represent elastic and plastic characteristic speeds; the outer arrows have slope

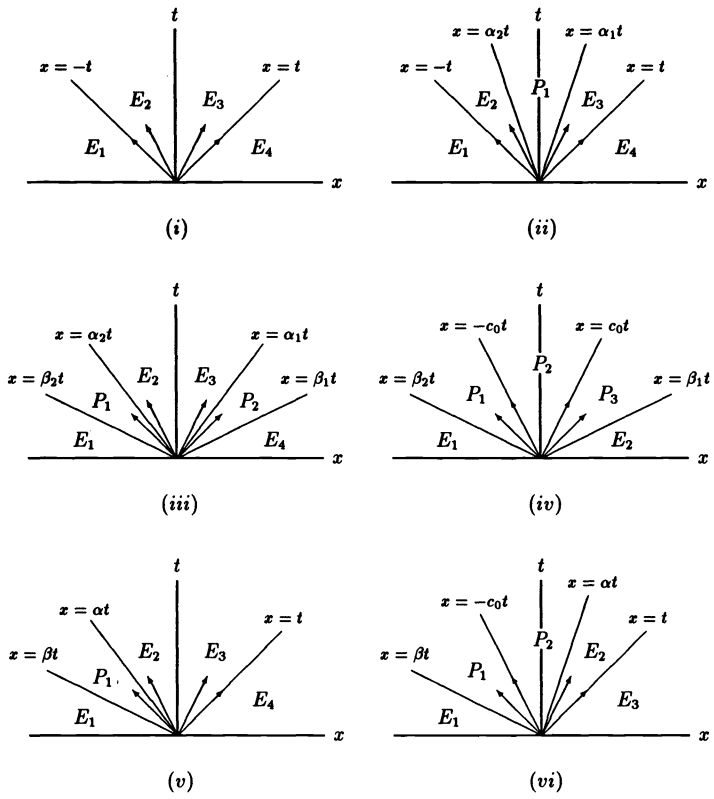


FIG. 1.1.

± 1 and the inner arrows have slope $\pm c_0$. The cases in Fig. 1.1 are numbered according to which case of (1.4) is satisfied.

The following is the main result of this section.

THEOREM 1.1. *Let $U \in \mathcal{U}$ be a solution of* (0.1) *with nondegenerate initial data* (0.2). *Then U is of the same form as $\bar{U}$ in a neighborhood of the origin.*

We need a few technical lemmas to prove Theorem 1.1.

LEMMA 1.2. *Let $U \in \mathcal{U}$ and suppose that, for some $i, \bar{\sigma}_i < \bar{\gamma}_i$ on a border of $\bar{T}_i$. Then $\sigma_i < \gamma_i$ on the interior of T_i in a neighborhood of the origin.*

The proof is easy and we omit it.

LEMMA 1.3. *Let $U \in \mathcal{U}$ satisfy* (0.1), (0.2) *and suppose that $\bar{T}_i$ shares a border $x = \eta t$ with $\bar{T}_j$. Then*

$$(\eta^2 - \lambda_i^2)\partial_x \bar{v}_i = (\eta^2 - \lambda_j^2)\partial_x \bar{v}_j,$$

where $\lambda_k = c_0$ if $\bar{U}_k$ satisfies (1.3(a)) (*exclusively*) *and $\lambda_k = 1$ if $\bar{U}_k$ satisfies* (1.3(b)) (*or both cases of* (1.3)).

Proof. Since $\eta_k'(0) = \eta$ for all $x = \eta_k(t)$ between T_i and T_j, (1.2) implies

$$(\partial_x \bar{U}_i - \partial_x \bar{U}_j)\eta + (\partial_t \bar{U}_i - \partial_t \bar{U}_j) = 0.$$

Combining this with (1.3) gives

$$\eta \partial_x \bar{v}_i - \eta \partial_x \bar{v}_j + \partial_x \bar{\sigma}_i - \partial_x \bar{\sigma}_j = 0,$$

$$\eta \partial_x \bar{\sigma}_i - \eta \partial_x \bar{\sigma}_j + \lambda_i^2 \partial_x \bar{v}_i - \lambda_j^2 \partial_x \bar{v}_j = 0.$$

Eliminating $\partial_x \bar{\sigma}_i$ and $\partial_x \bar{\sigma}_j$ gives the result. □

In [5] it is shown that the direction of a plastic-elastic interface determines whether the plastic state lies to the left or right of the interface. We will need a slight generalization of this result in the proof of Theorem 1.1.

It is easy to see from (0.3) that $U^+_{i,x}(t), U^+_{i,t}(t)$ satisfy one of the following:

$$\begin{aligned} &\text{(a)} \quad v^+_{i,t}(t) = \sigma^+_{i,x}(t), \\ &\qquad \sigma^+_{i,t}(t) = c(\sigma_i(\eta_i(t),t))^2 v^+_{i,x}(t), \\ &\qquad \gamma^+_{i,t}(t) = \sigma^+_{i,t}(t) \geq 0; \\ &\text{(b)} \quad v^+_{i,t}(t) = \sigma^+_{i,x}(t), \\ &\qquad \sigma^+_{i,t}(t) = v^+_{i,x}(t), \\ &\qquad \gamma^+_{i,t}(t) = 0. \end{aligned} \tag{1.5}$$

The same is true of $U^-_{i,x}(t), U^-_{i,t}(t)$. Let

$$\nu_i(t) = \frac{\eta_i'(t)^2 - c(\sigma_i(\eta_i(t),t))^2}{\eta_i'(t)(\eta_i'(t)^2 - 1)}.$$

Then we have the following lemma.

LEMMA 1.4. *Let $U \in \mathcal{U}$ satisfy* (0.1), (0.2) *and suppose that for some $t > 0$*

$$\begin{aligned} &\text{(i)} \quad U^+_{i,x}(t) \neq U^-_{i,x}(t) \quad \textit{or} \quad U^+_{i,t}(t) \neq U^-_{i,t}(t), \\ &\text{(ii)} \quad \eta_i'(t) \neq 0, 1, c(\sigma_i(\eta_i(t),t)). \end{aligned} \tag{1.6}$$

Then, if $\nu_i(t) < 0$ (resp., $\nu_i(t) > 0$), $U^+_{i,x}(t), U^+_{i,t}(t)$ satisfy (1.5(a)) (*resp.*, (1.5(b))) *and $U^-_{i,x}(t), U^-_{i,t}(t)$ satisfy* (1.5(b)) (*resp.*, (1.5(a))).

Proof. Assume $\nu_i(t) < 0$ (the case $\nu_i(t) > 0$ is similar). It is not hard to show from (1.1), (1.5), and (1.6) that $U^+_{i,x}(t), U^+_{i,t}(t)$ and $U^-_{i,x}(t), U^-_{i,t}(t)$ must satisfy different cases of (1.5). (This is essentially because singularities propagate along characteristics for first-order hyperbolic systems.) In particular, this means that

$$v^\pm_{i,x}(t) \neq 0 \tag{1.7}$$

(otherwise both cases of (1.5) would be satisfied simultaneously). We suppose that $U^+_{i,x}(t), U^+_{i,t}(t)$ satisfy (1.5(b)) and arrive at a contradiction. This implies that $\sigma_i < \gamma_i$ and $\sigma_{i-1} = \gamma_{i-1}$ in a neighborhood of $(\eta_i(t), t)$. Hence

$$\sigma^+_{i,x}(t) \leq \gamma^+_{i,x}(t) \tag{1.8}$$

and

$$\eta_i'(t)\sigma^+_{i,x}(t) + \sigma^+_{i,t}(t) = \eta_i'(t)\gamma^+_{i,x}(t) + \gamma^+_{i,t}(t). \tag{1.9}$$

Combining (1.5(b)), (1.7), (1.8), and (1.9), we have

$$\frac{v^+_{i,x}(t)}{\eta_i'(t)} = \gamma^+_{i,x}(t) - \sigma^+_{i,x}(t) > 0. \tag{1.10}$$

Now, using (1.1), (1.5), and arguing as in the proof of Lemma 1.4, we get

$$(\eta_i'(t)^2 - 1)v_{i,x}^+(t) = (\eta_i'(t)^2 - c(\sigma_i(\eta_i(t),t))^2)v_{i,x}^-(t). \tag{1.11}$$

Equations (1.10), (1.11) imply that $\nu_i(t)v_{i,x}^-(t) > 0$ and so $v_{i,x}^-(t) < 0$, which contradicts the fact that $U_{i,x}^-(t), U_{i,t}^-(t)$ satisfy (1.5(a)). This completes the proof. □

Proof of Theorem 1.1. Since $\bar{U}$ is the solution to the scale-invariant problem, the following facts follow from [5] and the strictness of the inequalities in (1.4).

(A) $\partial_x \bar{v} > 0$ on regions where $\bar{U}$ is plastic.

(B) Plastic-elastic interfaces of $\bar{U}$ are noncharacteristic.

(C) $\bar{\sigma} < \bar{\gamma}$ on the interiors of, and on boundaries between, regions where $\bar{U}$ is elastic.

Suppose $\bar{T}_i$ has nonempty interior; then it is a plastic or elastic region of $\bar{U}$. If plastic, then $\partial_t \bar{\gamma}_i = \partial_t \bar{\sigma}_i > 0$ by (A), which implies that $\partial_t \gamma_i$ is locally positive, and so U_i must be locally plastic. If elastic, then $\sigma_i < \gamma_i$ locally by (C) and Lemma 1.2, and so U_i is locally elastic.

We now show that no $\bar{T}_i$ has empty interior. Suppose otherwise and consider the following three cases.

(1) Suppose $\bar{T}_i$ separates elastic regions of $\bar{U}$, or $\bar{T}_i$ lies on the x-axis. Then (C) and Lemma 1.2 imply that U_i is locally elastic, but this implies that at least one boundary of T_i is not locally characteristic and separates elastic regions of U, contradicting (iv) of Definition 1.1.

(2) Suppose $\bar{T}_i$ separates elastic and plastic regions $\bar{T}_j, \bar{T}_k$. Let r be between j and k. Applying Lemma 1.3 to $\bar{U}_r, \bar{U}_k$, we have $\partial_x \bar{v}_r \neq 0$ because of (A), (B). This implies that U_r is either locally plastic or locally elastic. Since U_j is locally elastic and U_k is locally plastic, there must be some locally noncharacteristic $x = \eta_s(t)$ separating regions of the same type, again contradicting Definition 1.1 (iv).

(3) Suppose $\bar{T}_i$ separates plastic regions $\bar{T}_j, \bar{T}_k, j < k$. Then (A) implies that U_j and U_k are locally plastic. First assume $\eta_i'(0) \neq \pm c_0$. If $\eta_i'(0) \neq 0, \pm 1$ then (1.6) is locally satisfied along $x = \eta_{j+1}(t), \eta_k(t)$, and $\nu_{j+1}(t), \nu_k(t)$ have the same sign (locally). This contradicts Lemma 1.4. If $\eta_i'(0) = 0, \pm 1$ then there are either points on $x = \eta_{j+1}(t)$ where (1.6) is satisfied and $\nu_{j+1}(t) < 0$, or on $x = \eta_k(t)$ where (1.6) is satisfied and $\nu_k(t) > 0$. Again Lemma 1.4 is contradicted.

Now assume that $\eta_i'(0) = \pm c_0$. It is not hard to show that

$$0 < \min\{\partial_x \bar{v}_j, \partial_x \bar{v}_k\} \leq \lim_{t \to 0+} \frac{v(\eta_k(t),t) - v(\eta_{j+1}(t),t)}{\eta_k(t) - \eta_{j+1}(t)} \leq \sum_{j<r<k} \partial_x \bar{v}_r,$$

so $\partial_x \bar{v}_r > 0$ for some $j < r < k$. U_r must be locally plastic, otherwise Lemma 1.3 applied to $\bar{U}_r, \bar{U}_k$ implies that $\partial_x \bar{v}_r = 0$. Then there are either points on $x = \eta_r(t)$ where (1.6) is satisfied and $\nu_r(t) > 0$, or on $x = \eta_{r+1}(t)$ where (1.6) is satisfied and $\nu_{r+1}(t) < 0$, contradicting Lemma 1.4. This completes the proof of Theorem 1.1. □

2. Construction of the solutions. The following is our main result.

THEOREM 2.1. *If* (0.2) *is nondegenerate then there is a solution* $U \in \mathcal{U}$ *of* (0.1), (0.2). *Moreover,* U *is unique within* $\mathcal{U}$, *i.e., if* $\tilde{U} \in \mathcal{U}$ *is a solution of* (0.1), (0.2) *then* U *and* $\tilde{U}$ *coincide on the intersection of their domains.*

This is proven by constructing a solution for each case of (1.4) of corresponding form in Fig. 1.1. It is clear from the construction that it is the only solution of

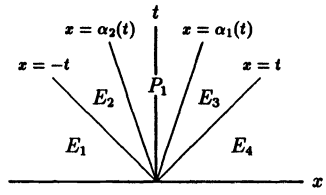


FIG. 2.1.

that form and that, locally, $\sigma < \gamma$ on elastic regions and $\partial_t \sigma > 0$ on plastic regions. Uniqueness then follows from Theorem 1.1. We will carry out the details of the proof only in Case (ii); the other cases are either straightforward or simplifications of Case (ii) (see [3]).

Case i. The solution is locally elastic and has the form shown in Fig. 1.1 (i), so U is found by simply following characteristics.

Case ii. Locally, U has the form shown in Fig. 2.1. U is smooth and plastic on P_1 and elastic on the E_i's, and α_1, α_2 are smooth functions such that

$$-c_0 < \alpha_2' < 0 < \alpha_1' < c_0. \tag{2.1}$$

Let $v = v_L, \sigma = \sigma_L, \gamma = \gamma_L$ on E_1 and $v = v_R, \sigma = \sigma_R, \gamma = \gamma_R$ on E_4. Then by following elastic characteristics we have

$$\begin{aligned} v_R(x,t) &= a_R x + b_R t + F(x,t), \\ \sigma_R(x,t) &= b_R x + a_R t + G(x,t), \\ \gamma_R(x,t) &= c_R x + h(x); \\ v_L(x,t) &= a_L x + b_L t + F(x,t), \\ \sigma_L(x,t) &= b_L x + a_L t + G(x,t), \\ \gamma_L(x,t) &= c_L x + h(x), \end{aligned} \tag{2.2}$$

where

$$\begin{aligned} F(x,t) &= \frac{1}{2}[f(x+t) + f(x-t) + g(x+t) - g(x-t)], \\ G(x,t) &= \frac{1}{2}[f(x+t) - f(x-t) + g(x+t) + g(x-t)] \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} \sigma + v &= \sigma_R + v_R, \quad \gamma = \gamma_R \quad \text{in } E_3, \\ \sigma - v &= \sigma_L - v_L, \quad \gamma = \gamma_L \quad \text{in } E_2. \end{aligned} \tag{2.4}$$

We know that $\sigma = \gamma$ on $x = \alpha_1(t), \alpha_2(t)$. Combining this with (2.4), we have the following boundary conditions on P_1.

$$\begin{aligned} v &= v_r, \quad \sigma = \sigma_r \quad \text{on } x = \alpha_1(t), \\ v &= v_l, \quad \sigma = \sigma_l \quad \text{on } x = \alpha_2(t), \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} v_r &= v_R + \sigma_R - \gamma_R, \qquad \sigma_r = \gamma_R, \\ v_l &= v_L - \sigma_L + \gamma_L, \qquad \sigma_l = \gamma_L. \end{aligned} \tag{2.6}$$

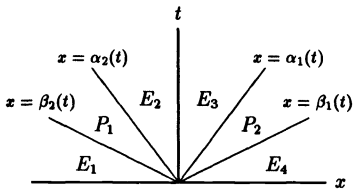


FIG. 2.2.

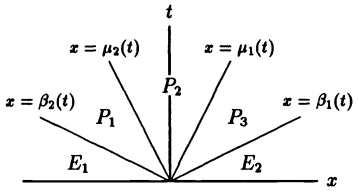


FIG. 2.3.

Notice that neither of the plastic-elastic interfaces, $x = \alpha_1(t)$ and $x = \alpha_2(t)$, can be determined from information on only one side. We find them by solving a free boundary problem on P_1 with boundary data given by (2.5), (2.6) and with the plastic equations

$$\partial_t v = \partial_x \sigma, \qquad \partial_t \sigma = c(\sigma)^2 \partial_x v. \tag{2.7}$$

If α_1 and α_2 were a priori known, there would be two more boundary conditions than necessary to solve this problem, but since α_1, α_2 are unknown it is reasonable to expect a unique solution. We solve (2.5), (2.7), satisfying (2.1) in §4.

Case iii. Locally, U has the form shown in Fig. 2.2. Also, $\beta_2' < -1 < \alpha_2' < -c_0$ and $c_0 < \alpha_1' < 1 < \beta_1'$.

The solution in E_1, E_4 is as in Case ii. We then have implicit equations $\sigma_R = \gamma_R$ and $\sigma_L = \gamma_L$ for the curves $x = \beta_1(t)$ and $x = \beta_2(t)$, resp., which can be shown to be locally solvable using (1.4(iii)) and (2.2). Since both σ and v are known on $x = \beta_1(t)$ and $x = \beta_2(t)$, we now have a Cauchy problem for the plastic equations on P_1 and P_2, which can be locally solved using the method of [6, Chap. 2]. Then both σ and v are known on the curves $x = \alpha_1(t)$ and $x = \alpha_2(t)$, leading to a free boundary problem, similar to that of Case ii, for the elastic equations on $E_2 \cup E_3$. Since the equations are constant coefficient instead of quasi-linear, this problem can be solved by a simplification of the method used in Case ii. We will remark on the differences in §4.

Case iv. Locally, U has the form shown in Fig. 2.3. Also, $\beta_2' < -1, \beta_1' > 1$, and $\mu_1'(0) = c_0, \mu_2'(0) = -c_0$.

The solution in E_1, E_2, P_1, P_3, and the curves $x = \beta_1(t), \beta_2(t)$ are found as in Case iii. Using the method of Remark 4.2 in [6, Chap. 2], we can maximize the domain of the solution to the Cauchy problem on P_1 and P_3, thereby finding the characteristic curves $x = \mu_1(t)$ and $x = \mu_2(t)$. The problem is thus reduced to a Goursat problem on P_2, which can be solved using the method of [6, Chap. 3].

Case v. Locally, U has the form shown in Fig. 2.4. Also, $\beta' < -1 < \alpha' < -c_0$.

The solution in E_1, E_4, P_1 and the curve $x = \beta(t)$ are found as in Case iii. We then get an implicit equation for the curve $x = \alpha(t)$ by setting the solution for $\sigma + v$ in P_1 equal to $\sigma_R + v_R$ (following the left-moving characteristic from E_4). The remaining unknowns on E_2 and E_3 can then be found by following characteristics.

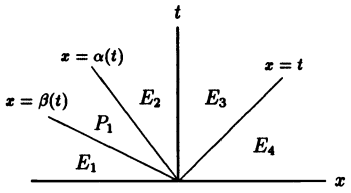


FIG. 2.4.

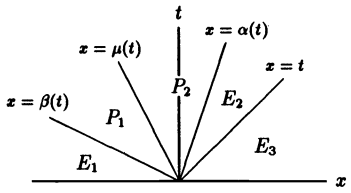


FIG. 2.5.

Case vi. Locally, U has the form shown in Fig. 2.5. Also, $\beta' < -1, 0 < \alpha' < c_0$, and $\mu'(0) = -c_0$.

The solution in E_1, E_3, P_1 and the curves $x = \beta(t)$ and $x = \mu(t)$ are found as in Case iv. We then have a free boundary problem on P_2 with free boundary $x = \alpha(t)$. This problem can be solved by a simplification of the method used in Case ii, and we will give the argument in a remark at the end of §4.

3. Solution of a fixed boundary problem. The purpose of this section is to solve a Goursat problem for (3.1) which will be needed to solve the free boundary problem derived in Case ii of §2. In Theorem 3.1 we show that this Goursat problem has a solution and derive estimates which will be useful in proving convergence of the interfaces to a solution of (2.5), (2.7). First we make a convenient change of variables and coordinates.

Define

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

by

$$u_1 = \sigma + c_0 v, \qquad u_2 = \sigma - c_0 v, \tag{3.1}$$

and

$$y = \frac{1}{2}(t + x/c_0), \qquad z = \frac{1}{2}(t - x/c_0). \tag{3.2}$$

Let $z = \rho_1(y)$ and $y = \rho_2(z)$ be equivalent to $x = \alpha_1(t), \alpha_2(t)$. Then the boundary conditions (2.5) become

$$\begin{aligned} u &= u^r \quad \text{on } z = \rho_1(y), \\ u &= u^l \quad \text{on } y = \rho_2(z), \end{aligned} \tag{3.3}$$

where $0 < \rho_1'(0), \rho_2'(0) < 1$ because of (2.1), and

$$\begin{aligned} u_1^r &= \sigma_r + c_0 v_r, \qquad u_2^r = \sigma_r - c_0 v_r, \\ u_1^l &= \sigma_l + c_0 v_l, \qquad u_2^l = \sigma_l - c_0 v_l. \end{aligned} \tag{3.4}$$

We write the transformed plastic equations (2.7) in characteristic form:

$$\Lambda_1(u)A(u)\partial_y u + \Lambda_2(u)A(u)\partial_z u = 0, \tag{3.5}$$

where

$$\Lambda_1(u) = \begin{bmatrix} \lambda(u) & 0 \\ 0 & 1 \end{bmatrix}, \qquad \Lambda_2(u) = \begin{bmatrix} 1 & 0 \\ 0 & \lambda(u) \end{bmatrix},$$
$$A(u) = \begin{bmatrix} 1 & \lambda(u) \\ \lambda(u) & 1 \end{bmatrix}, \qquad \lambda(u) = \frac{c_0 - c(\sigma)}{c_0 + c(\sigma)}.$$

We will solve (3.3), (3.4), (3.5) in §4, but first we consider the following boundary value problem for (3.5). Let $\rho_1, \rho_2 \in C^1[0,\infty)$ satisfy $\rho_i(0) = 0, 0 < \rho_1'(0), \rho_2'(0) < 1$. Suppose we have boundary conditions

$$u_1 = \psi_1 \quad \text{on } z = \rho_1(y), \qquad u_2 = \psi_2 \quad \text{on } y = \rho_2(z), \tag{3.6}$$

where $\psi_i(y,z)$ is smooth and zero at the origin. Choose $\kappa > 0$ such that $\kappa < \rho_1'(0), \rho_2'(0)$ and define a region $R(\delta)$ in the yz-plane as follows. Let $\tau_i(\delta)$ be defined by

$$\tau_i(\delta) - \delta = \kappa\delta - \kappa\rho_i(\tau_i(\delta)), \qquad i = 1,2. \tag{3.7}$$

Let

$$R(\delta) = \{(y,z) : \kappa z \le y \le \pm\kappa(z - \rho_1(\tau_1(\delta))) + \tau_1(\delta),$$
$$\kappa y \le z \le \pm\kappa(y - \rho_2(\tau_2(\delta))) + \tau_2(\delta)\}$$

(see Fig. 3.1). Then we have the following theorem.

THEOREM 3.1. *There is some $\delta_* > 0$ such that* (3.5), (3.6) *has a unique smooth solution on $R(\delta_*)$.*

Imitating [6, Chap. 2], we prove this theorem by means of an iterative process which solves an analogous linear problem at each step. In [6], more specialized boundary conditions than (3.6) are assumed for the linear problem, resulting in a possibly larger domain of existence. However, this does not have a significant effect on the domain of the solution to the quasi-linear problem, and the proof of Theorem 3.1 is simplified by using (3.6).

Consider the linear system

$$\Lambda_1(y,z)A(y,z)\partial_y u + \Lambda_2(y,z)A(y,z)\partial_z u = 0, \tag{3.8}$$

where Λ_i, A are defined as before with $\tilde{\lambda}(y,z)$ replacing $\lambda(u)$, and $\tilde{\lambda}$ is smooth and zero at the origin.

LEMMA 3.1. *There is some $\delta_0 > 0$ such that* (3.8), (3.6) *has a unique smooth solution on $R(\delta_0)$.*

Proof. First we integrate along characteristics and put (3.8), (3.6) in an integral form. Let $\xi = f(\zeta; y, z), \zeta = g(\xi; y, z)$ be defined by

$$\begin{aligned} &\frac{df}{d\zeta}(\zeta; y, z) = \tilde{\lambda}(f(\zeta; y, z), \zeta), \\ &\frac{dg}{d\xi}(\xi; y, z) = \tilde{\lambda}(\xi, g(\xi; y, z)), \\ &f(z; y, z) = y, \qquad g(y; y, z) = z \end{aligned} \tag{3.9}$$

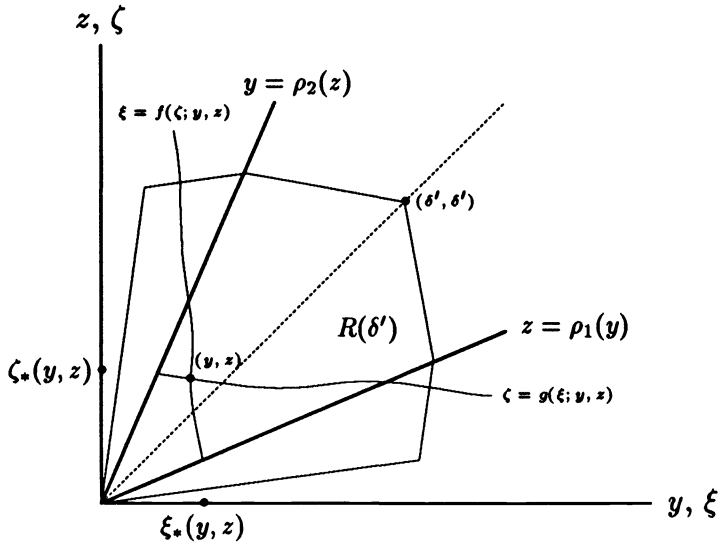


FIG. 3.1.

(see Fig. 3.1). Let $\xi_*(y,z), \zeta_*(y,z)$ be smooth functions satisfying

$$f(\rho_1(\xi_*); y, z) = \xi_*, \qquad g(\rho_2(\zeta_*); y, z) = \zeta_*. \tag{3.10}$$

Choose $\delta' > 0$ such that

$$|\tilde{\lambda}| < \kappa \le \rho_1', \rho_2' \le 1 \quad \text{on } R(\delta'). \tag{3.11}$$

Equation (3.11) implies that $R(\delta')$ is a domain of determinacy for (3.8), (3.6), so we can write (3.8), (3.6) in the following integral form on $R(\delta')$:

$$u(y,z) = A(y,z)^{-1} \begin{pmatrix} (\psi_1 + \tilde{\lambda} u_2)(\xi_*, \rho_1(\xi_*)) + \int_{\rho_1(\xi_*)}^{z} (u_2 \partial_\zeta \tilde{\lambda})(f, \zeta)\, d\zeta \\ (\psi_2 + \tilde{\lambda} u_1)(\rho_2(\zeta_*), \zeta_*) + \int_{\rho_2(\zeta_*)}^{y} (u_1 \partial_\xi \tilde{\lambda})(\xi, g)\, d\xi \end{pmatrix}, \tag{3.12}$$

where $f = f(\zeta; y, z), g = g(\xi; y, z), \partial_\zeta = \tilde{\lambda}\partial_y + \partial_z, \partial_\xi = \tilde{\lambda}\partial_z + \partial_y$.

We can write (3.12) as a fixed point problem, $u = Tu$, where T is the integral operator on the right side. It is easy to see that, for $u, \bar{u} \in C^1(R(\delta'))$,

$$\|T\bar{u} - Tu\|_\delta \le A_1 \delta \|\bar{u} - u\|_\delta \tag{3.13}$$

for $\delta \le \delta'$, where $\|\cdot\|_\delta = \sup_{R(\delta)} |\cdot|$ and A_1 depends on $(1 - \|\tilde{\lambda}\|_{\delta'})^{-1}, \|\nabla\tilde{\lambda}\|_{\delta'}$. As in the proof of Theorem 1.1 of [6, Chap. 1], we can show that $T : C^1(R(\delta')) \to C^1(R(\delta'))$. Since we only assume that $\tilde{\lambda}$ is C^1, we see from (3.12) that we cannot directly differentiate Tu with respect to y and z. The proof involves approximating the derivatives with difference quotients, integrating by parts where necessary, and then passing to a limit. This leads to an expression for ∇Tu (which we omit). From this expression, it is not hard to show that, for $u, \bar{u} \in C^1(R(\delta'))$,

$$\|\nabla T\bar{u} - \nabla Tu\|_\delta \le A_2 \delta \|\nabla\bar{u} - \nabla u\|_\delta \tag{3.14}$$

for $\delta \leq \delta'$, where A_2 depends on $(1 - \|\tilde{\lambda}\|_{\delta'})^{-1}, \|\nabla\tilde{\lambda}\|_{\delta'}$. Choose $\delta_0 \in (0, \delta']$ such that $A_i\delta_0 < 1, i = 1, 2$, and consider the sequence

$$u^{(n+1)} = Tu^{(n)}, \qquad u^{(0)} = 0.$$

By (3.13), (3.14), $u^{(n)}$ converges uniformly to a unique solution $u \in C^1(R(\delta_0))$ of (3.12), proving the lemma. □

We will need estimates of the solution to the linear problem, (3.8), (3.6), in order to solve the quasi-linear problem, (3.5), (3.6). Let $|x| = \sup_{1\leq i\leq n} |x_i|$ for $x \in \mathbf{R}^n$. Define the modulus of continuity of $f : \mathbf{R}^n \to \mathbf{R}^m$ by

$$\omega(\varepsilon, \eta|f) = \sup_{\substack{|\bar{x}-x|\leq\eta \\ |\bar{x}|,|x|\leq\varepsilon}} |f(\bar{x}) - f(x)|.$$

For a set of functions $\{f_1, \ldots, f_k\}$, let $\omega(\varepsilon, \eta|\{f_k\}) = \max_{1\leq i\leq k} \omega(\varepsilon, \eta|f_i)$.

The modulus of continuity has the following properties.

$$\omega(\varepsilon, \eta|fg) \leq |f|_\varepsilon\omega(\varepsilon, \eta|g) + |g|_\varepsilon\omega(\varepsilon, \eta|f), \tag{3.15}$$

where $|\cdot|_\varepsilon = \sup_{|x|\leq\varepsilon} |\cdot|$.

$$\omega(\varepsilon, \eta|f/g) \leq |f|_\varepsilon|1/g|_\varepsilon^2\omega(\varepsilon, \eta|g) + |1/g|_\varepsilon\omega(\varepsilon, \eta|f). \tag{3.16}$$

$$\omega(\varepsilon, C\eta|f) \leq ([C] + 1)\omega(\varepsilon, \eta|f), \tag{3.17}$$

where $[\cdot]$ is the greatest integer function.

$$\omega(\varepsilon, \eta|f \circ g) \leq \omega(|g|_\varepsilon, \eta|\nabla g|_\varepsilon|f). \tag{3.18}$$

Define Ω, the modulus of continuity on $R(\delta)$, by

$$\Omega(\delta, \eta|h(y, z)) = \sup_{\substack{|\bar{y}-y|,|\bar{z}-z|\leq\eta \\ (y,z),(\bar{y},\bar{z})\in R(\delta)}} |h(\bar{y}, \bar{z}) - h(y, z)|.$$

Ω satisfies (3.15)–(3.18) with $\|\cdot\|_\delta$ in place of $|\cdot|_\varepsilon$ where appropriate. Also, if

$$h(y, z) = \int_{a(y,z)}^{b(y,z)} g(s; y, z)\, ds,$$

then

$$\Omega(\delta, \eta|h) \leq \|g\|_\delta[\Omega(\delta, \eta|a) + \Omega(\delta, \eta|b)] + \|b - a\|_\delta \Big\| \sup_{a\leq s\leq b} \Omega(\delta, \eta|g(s; y, z))\Big\|_\delta, \tag{3.19}$$

where the modulus in the supremum is taken with respect to y, z.

Define functions

$$\psi_*(\eta) = \Omega(\delta_0, \eta|\nabla\psi), \quad \tilde{\lambda}_*(\eta) = \Omega(\delta_0, \eta|\nabla\tilde{\lambda}), \quad \rho_*(\eta) = \max_{i=1,2}\{\omega(\tau_i(\delta_0), \eta|\rho_i')\},$$

where $\psi = (\psi_1, \psi_2)$.

LEMMA 3.2. *Let u be the solution to* (3.8), (3.6) *on* $R(\delta_0)$. *Then there is some* $\delta_1 \in (0, \delta_0]$, *depending on* $(1 - \|\tilde{\lambda}\|_{\delta_0})^{-1}, \|\nabla\tilde{\lambda}\|_{\delta_0}$, *such that*

(a) $\|\nabla u\|_\delta \le (2 + K_1\delta)\|\nabla\psi\|_\delta$ *for* $\delta \le \delta_1$, *where* K_1 *depends on* $(1 - \|\tilde{\lambda}\|_{\delta_0})^{-1}$, $\|\nabla\tilde{\lambda}\|_{\delta_0}$.

(b) $\Omega(\delta, \eta|\nabla u) \le K_2(\psi_*(\eta) + \rho_*(\eta) + \eta + \delta\tilde{\lambda}_*(\eta))$ *for* $\delta \le \delta_1$, *where* K_2 *depends on* $(1 - \|\tilde{\lambda}\|_{\delta_0})^{-1}, \|\nabla\psi\|_{\delta_0}, \|\nabla\tilde{\lambda}\|_{\delta_0}$.

Proof. Let $p = \partial_y u, q = \partial_z u$. Imitating the method of [6, §1, Chap. 2], we can derive integral equations for p and q. We formally differentiate the system (3.8), then argue as in the proof of Lemma 3.1 to obtain integral equations

$$
\begin{aligned}
p(y,z) &= A(y,z)^{-1}\begin{pmatrix} a_1(\xi_*, \rho_1(\xi_*)) + \int_{\rho_1(\xi_*)}^{z}(p_2\partial_\zeta\tilde{\lambda} - b_1\partial_y\tilde{\lambda})(f,\zeta)\,d\zeta \\ -(\tilde{\lambda}a_2)(\rho_2(\zeta_*),\zeta_*) + \int_{\rho_2(\zeta_*)}^{y}(p_1\partial_\xi\tilde{\lambda} - b_2\partial_y\tilde{\lambda})(\xi,g)\,d\xi \end{pmatrix}, \\
q(y,z) &= A(y,z)^{-1}\begin{pmatrix} -(\tilde{\lambda}a_1)(\xi_*, \rho_1(\xi_*)) + \int_{\rho_1(\xi_*)}^{z}(q_2\partial_\zeta\tilde{\lambda} - b_1\partial_z\tilde{\lambda})(f,\zeta)\,d\zeta \\ a_2(\rho_2(\zeta_*),\zeta_*) + \int_{\rho_2(\zeta_*)}^{y}(q_1\partial_\xi\tilde{\lambda} - b_2\partial_z\tilde{\lambda})(\xi,g)\,d\xi \end{pmatrix},
\end{aligned} \tag{3.20}
$$

where $b_1 = q_2 + p_1 + 2\tilde{\lambda}p_2, b_2 = q_2 + p_1 + 2\tilde{\lambda}q_1$, and

$$
a_1 = \frac{\partial_{(1)}\psi_1 + \tilde{\lambda}\partial_{(1)}u_2}{1 - \rho_1'\tilde{\lambda}}, \qquad a_2 = \frac{\partial_{(2)}\psi_2 + \tilde{\lambda}\partial_{(2)}u_1}{1 - \rho_2'\tilde{\lambda}},
$$

where $\partial_{(1)} = \rho_1'\partial_z + \partial_y, \partial_{(2)} = \rho_2'\partial_y + \partial_z$. This formal process can be considered valid since we can first obtain the system of integral equations satisfied by difference quotients, integrate by parts where necessary, and then pass to a limit.

From (3.20) we can see that

$$
\|\nabla u\|_\delta \le (2 + K_3\delta)\|\nabla\psi\|_\delta + K_3\delta\|\nabla u\|_\delta
$$

for $\delta \le \delta_0$, where K_3 depends on $(1 - \|\tilde{\lambda}\|_{\delta_0})^{-1}, \|\nabla\tilde{\lambda}\|_{\delta_0}$. Part (a) of the lemma then follows.

Imitating the proof of Lemma 3.3 in [6, Chap. 1], we can use (3.9) to show that

$$
\|\partial_y f\|_\delta, \|\partial_z g\|_\delta \le 1 + K_4\delta; \qquad \|\partial_z f\|_\delta, \|\partial_y g\|_\delta \le K_4\delta \tag{3.21}
$$

for $\delta \le \delta_0$, where K_4 depends on $\|\nabla\tilde{\lambda}\|_{\delta_0}$. From (3.10) we can show that

$$
\nabla\xi_* = \frac{\nabla_{(y,z)}f(\rho_1(\xi_*);\cdot,\cdot)}{1 - \rho_1'(\xi_*)\tilde{\lambda}(\xi_*, \rho_1(\xi_*))}, \qquad \nabla\zeta_* = \frac{\nabla_{(y,z)}g(\rho_2(\zeta_*);\cdot,\cdot)}{1 - \rho_2'(\zeta_*)\tilde{\lambda}(\rho_2(\zeta_*),\zeta_*)}. \tag{3.22}
$$

Combining (3.21), (3.22) we have

$$
\|\nabla\xi_*\|_\delta, \|\nabla\zeta_*\|_\delta \le 1 + K_5\delta \tag{3.23}
$$

for $\delta \le \delta_0$, where K_5 depends on $(1 - \|\tilde{\lambda}\|_{\delta_0})^{-1}, \|\nabla\tilde{\lambda}\|_{\delta_0}$. Now, applying (3.15)–(3.19) to (3.20), and using (3.23), we have

$$
\Omega(\delta, \eta|\nabla u) \le K_6(\psi_*(\eta) + \rho_*(\eta) + \eta + \delta\tilde{\lambda}_*(\eta) + \delta\Omega(\delta, \eta|\nabla u))
$$

for $\delta \le \delta_0$, where K_6 depends on $(1 - \|\tilde{\lambda}\|_{\delta_0})^{-1}, \|\nabla u\|_{\delta_0}, \|\nabla\psi\|_{\delta_0}, \|\nabla\tilde{\lambda}\|_{\delta_0}$. Part (b) then follows from this and part (a). □

Proof of Theorem 3.1. Choose a constant $\Omega_1 > 2|\nabla\psi(0,0)|$ and $\varepsilon > 0$ such that

$$|\lambda(u)| < \kappa \quad \text{for } |u| \le \varepsilon. \tag{3.24}$$

Let $\Sigma(\delta) = \{u \in C^1(R(\delta)) : \|u\|_\delta \le \varepsilon, \|\nabla u\|_\delta \le \Omega_1\}$. Define an operator $\tilde{u} = Qu$ by letting $\tilde{u}$ be the solution to

$$\Lambda_1(u)A(u)\partial_y\tilde{u} + \Lambda_2(u)A(u)\partial_z\tilde{u} = 0 \tag{3.25}$$

with boundary conditions

$$\tilde{u}_1 = \psi_1 \quad \text{on } z = \rho_1(y), \qquad \tilde{u}_2 = \psi_2 \quad \text{on } y = \rho_2(z). \tag{3.26}$$

By the proof of Lemma 3.1, there is some $\delta_0 > 0$ such that (3.11) is satisfied, with $\tilde{\lambda} = \lambda(u)$, for any $u \in \Sigma(\delta_0)$ (because of (3.24)) and $Q : \Sigma(\delta) \to C^1(R(\delta))$ for $\delta \le \delta_0$. By Lemma 3.2(a), there is some $\delta_1 \in (0, \delta_0]$, depending on $(1 - |\lambda|_\varepsilon)^{-1}, \Omega_1, |\nabla_u\lambda|_\varepsilon$, such that

$$\|\nabla\tilde{u}\|_\delta \le (2 + \tilde{K}_1\delta)\|\nabla\psi\|_\delta \tag{3.27}$$

for all $u \in \Sigma(\delta)$, for $\delta \le \delta_1$, which implies

$$\|\tilde{u}\|_\delta \le \delta(2 + \tilde{K}_1\delta)\|\nabla\psi\|_\delta \tag{3.28}$$

for $\delta \le \delta_1$, where $\tilde{K}_1$ depends on $(1 - |\lambda|_\varepsilon)^{-1}, \Omega_1, |\nabla_u\lambda|_\varepsilon$. By (3.27), (3.28) there is some $\delta_2 \in (0, \delta_1]$ such that if $u \in \Sigma(\delta)$ then

$$\|\tilde{u}\|_\delta \le \varepsilon, \qquad \|\nabla\tilde{u}\|_\delta < \Omega_1 \tag{3.29}$$

for $\delta \le \delta_2$. Hence $Q : \Sigma(\delta) \to \Sigma(\delta)$ for $\delta \le \delta_2$.

Now, by Lemma 3.2(b), there is some $\delta_3 \in (0, \delta_2]$ such that

$$\Omega(\delta, \eta|\nabla\tilde{u}) \le \tilde{K}_2(\psi_*(\eta) + \rho_*(\eta) + \eta + \delta\Omega(\delta, \eta|\nabla u) + \delta\lambda_*(\eta)) \tag{3.30}$$

for all $u \in \Sigma(\delta)$, for $\delta \le \delta_3$, where $\lambda_* = \omega(\varepsilon, \eta|\nabla_u\lambda)$ and $\tilde{K}_2$ depends on $(1 - |\lambda|_\varepsilon)^{-1}, \Omega_1, |\nabla_u\lambda|_\varepsilon$. Let $\Omega_2(\eta) = 2\tilde{K}_2(\psi_*(\eta) + \rho_*(\eta) + \eta + \lambda_*(\eta))$ and

$$\Sigma_*(\delta) = \{u \in \Sigma(\delta) : \Omega(\delta, \eta|\nabla u) \le \Omega_2(\eta)\}.$$

Then (3.30) implies that there is some $\delta_4 \in (0, \delta_3]$ such that $Q : \Sigma_*(\delta) \to \Sigma_*(\delta)$ for $\delta \le \delta_4$.

We now show that Q is a contraction on some $\Sigma_*(\delta)$. Let $u^{(1)}, u^{(2)} \in \Sigma_*(\delta_4)$ and $\tilde{u}^{(i)} = Qu^{(i)}, i = 1, 2$. Let $u^* = u^{(2)} - u^{(1)}, \tilde{u}^* = \tilde{u}^{(2)} - \tilde{u}^{(1)}$. Clearly, $\tilde{u}^*$ satisfies

$$\Lambda_1(u^{(2)})A(u^{(2)})\partial_y\tilde{u}^* + \Lambda_2(u^{(2)})A(u^{(2)})\partial_z\tilde{u}^* = H(y, z), \tag{3.31}$$

where

$$\begin{aligned} H(y,z) = {} & (\Lambda_1(u^{(1)})A(u^{(1)}) - \Lambda_1(u^{(2)})A(u^{(2)}))\partial_y\tilde{u}^{(1)} \\ & + (\Lambda_2(u^{(1)})A(u^{(1)}) - \Lambda_2(u^{(2)})A(u^{(2)}))\partial_z\tilde{u}^{(1)} \end{aligned}$$

and

$$\tilde{u}_1^* = 0 \quad \text{on } z = \rho_1(y), \qquad \tilde{u}_2^* = 0 \quad \text{on } y = \rho_2(z). \tag{3.32}$$

We can put (3.31), (3.32) in an integral form similar to (3.12) with $\tilde{\lambda} = \lambda(u^{(2)}), \psi = 0$, and with H added to the integrand. From this integral equation, we can show that

$$\|\tilde{u}^*\|_\delta \le C_0\delta(\|H\|_\delta + \|\tilde{u}^*\|_\delta) \tag{3.33}$$

for $\delta \le \delta_4$, where C_0 depends on $(1 - |\lambda|_\varepsilon)^{-1}, \Omega_1, |\nabla_u\lambda|_\varepsilon$. It is not hard to show that

$$\|H\|_\delta \le C_1\Omega_1\|u^*\|_\delta \tag{3.34}$$

for $\delta \le \delta_4$, where C_1 depends on $|\nabla_u\lambda|_\varepsilon$. Combining (3.33), (3.34), we see that there is some $\delta_* \in (0, \delta_4]$ such that Q is a contraction on $\Sigma_*(\delta_*)$. It is not hard to show that $\Sigma_*(\delta_*)$ is complete in the supremum norm, and so (3.5), (3.6) has a unique solution in $\Sigma_*(\delta_*)$.

Finally, we show this solution is unique in $C^1(R(\delta_*))$. Let $u^{(1)} \in \Sigma_*(\delta_*), u^{(2)} \in C^1(R(\delta_*))$ be solutions. Let $u^* = u^{(2)} - u^{(1)}$, so u^* satisfies (3.31), (3.32) with $u^{(i)}, u^*$ in place of $\tilde{u}^{(i)}, \tilde{u}^*$. Let $\bar{\delta} \le \delta_*$ be maximal such that $\|\nabla u^{(2)}\|_{\bar{\delta}} \le \Omega_1$. Then the above argument shows that $u^* = 0$ on $R(\bar{\delta})$. It is not hard to see from (3.20) that

$$\nabla u^{(2)}(0,0) = \nabla u^{(1)}(0,0) = \begin{bmatrix} \partial_{(1)}\psi_1(0,0) & 0 \\ 0 & \partial_{(2)}\psi_2(0,0) \end{bmatrix}, \tag{3.35}$$

which implies that $\bar{\delta} > 0$ since $|\nabla u^{(2)}(0,0)| < \Omega_1$. Hence $\|\nabla u^{(2)}\|_{\bar{\delta}} < \Omega_1$ by (3.29), and so $\bar{\delta} = \delta_*$. This completes the proof of Theorem 3.1. □

4. Solution of the free boundary problem. In this section we solve the free boundary problem (3.3), (3.5) by means of an iterative process described as follows. Given approximations for the free boundaries $z = \rho_1(y)$ and $y = \rho_2(z)$, we solve a Goursat problem of the type in §3, using only one of the given boundary conditions on each side. We then obtain new approximate boundaries by correcting the boundaries to reduce the errors in the neglected boundary conditions. We show that this process converges to a solution. One difficulty which arises is a shrinking of the domain of the approximate boundaries with each iteration.

By Lemma 1.1 and [5], there exist unique $\rho_1^0, \rho_2^0 \in (0,1)$ such that $\rho_1'(0) = \rho_1^0$ and $\rho_2'(0) = \rho_2^0$ for any solution of (3.3), (3.5). Choose $\kappa > 0$ such that $\kappa < \rho_1^0, \rho_2^0$. Let $\rho \in C^1[0,\infty) \times C^1[0,\infty), \rho(0) = (0,0), \rho'(0) = (\rho_1^0, \rho_2^0)$. By Theorem 3.1, there is a unique smooth solution $u(\rho)$ to (3.5), on some $R(\delta)$, satisfying boundary conditions

$$u_1(\rho) = u_1^r \quad \text{on } z = \rho_1(y), \qquad u_2(\rho) = u_2^l \quad \text{on } y = \rho_2(z), \tag{4.1}$$

where u^r, u^l are as in (3.4). Define $S(\rho) = \tilde{\rho}$, where

$$\begin{aligned} u_2^r(y, \tilde{\rho}_1(y)) &= u_2(\rho)(y, \tilde{\rho}_1(y)), \\ u_1^l(\tilde{\rho}_2(z), z) &= u_1(\rho)(\tilde{\rho}_2(z), z). \end{aligned} \tag{4.2}$$

Solving (3.3), (3.5) is now equivalent to finding a fixed point of S.

Remark. The free boundary problem for Case iii can be written the same way by changing variables as in (3.1), (3.2), replacing c_0 by 1. The arguments that follow will also apply to Case iii, though the estimates are greatly simplified since the elastic equations have constant coefficients (see [3]).

We claim that repeated applications of S will converge to a fixed point of S, and offer the following motivation. By Lemma 1.1, we can find ρ_1^0 and ρ_2^0 by solving the

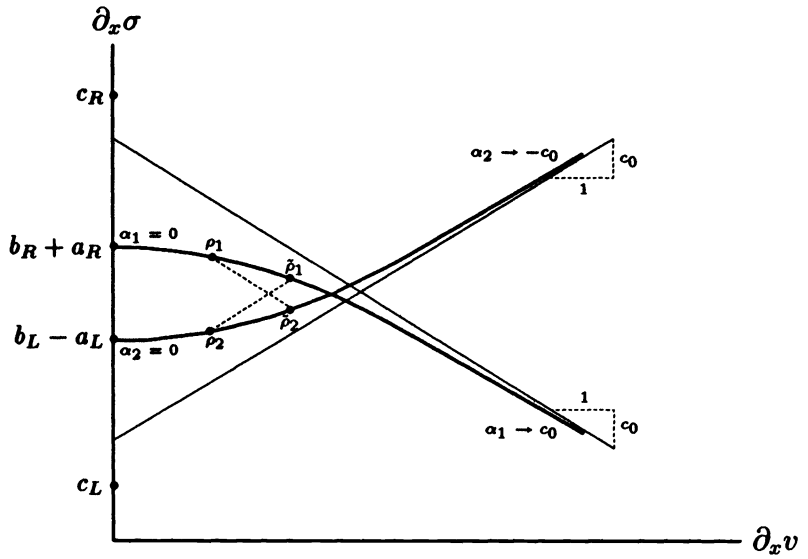


FIG. 4.1.

scale-invariant problem given by (0.1), (0.2) with $c(\sigma) \equiv c_0$ and $f, g, h \equiv 0$. In [5] this problem is solved by plotting possible plastic states (for P_1) in the $\partial_x v, \partial_x \sigma$-plane given α_1 (the speed of the right interface) and then doing the same given α_2 (the speed of the left interface). These wave curves have a unique point of intersection (see Fig. 4.1) which determines α_1 and α_2. Suppose we attempt to find this solution using the iterative scheme defined by (4.2). Since all functions involved are now linear in y and z, (4.2) reduces to a pair of algebraic equations which can be solved for constants $\tilde{\rho}_1$ and $\tilde{\rho}_2$ (which represent approximate slopes of the plastic-elastic interfaces in yz-space) in terms of constants ρ_1 and ρ_2. Figure 4.1 illustrates the relationship between these.

It is clear from the figure that this process is a contraction in $\partial_x v, \partial_x \sigma$-space, although it may not be a contraction in ρ-space, since the wave curves are parameterized differently in ρ_1 and ρ_2, and $\tilde{\rho}_2$ depends on ρ_1 and $\tilde{\rho}_1$ on ρ_2. However, the square of this process is a contraction in ρ-space near the solution since $\tilde{\tilde{\rho}}_i$ depends only on ρ_i. Returning to the general problem, we expect that, near the origin and for ρ_i' near ρ_i^0, the first-order effects of S will be like the linearized version. A similar argument can be given in Case iii; the size of the contraction differs, but the figure is qualitatively the same (see [3]).

We now find a fixed point of S. Let $F = u_2(\rho) - u_2^r, G = u_1(\rho) - u_1^l$. It is straightforward but tedious, using (2.2), (2.6), (3.1), (3.2), (3.4), and (3.35), to verify that $\partial_z F$ and $\partial_y G$ are nonzero at the origin, implying that $\tilde{\rho}$ exists locally. From now on we will use $R(\rho, \delta), \tau_i(\rho, \delta)$ to denote $R(\delta), \tau_i(\delta), i = 1, 2$, defined in §3. Define

$$M_0(\delta) = \{\rho \in C^1[0, \tau_1(\rho, \delta)] \times C^1[0, \tau_2(\rho, \delta)] : \rho(0) = (0, 0),$$
$$\rho'(0) = (\rho_1^0, \rho_2^0), \kappa \le \rho_1', \rho_2' \le 1\}.$$

It can be seen from the proof of Theorem 3.1 that there exist $\delta_1, \varepsilon, \Omega_1 > 0$ such that $u(\rho)$ exists on $R(\rho, \delta)$ for all $\rho \in M_0(\delta)$ and

$$\|u(\rho)\|_\delta \le \varepsilon, \qquad \|\nabla u(\rho)\|_\delta \le \Omega_1 \tag{4.3}$$

for $\delta \le \delta_1$, where $\|\cdot\|_\delta = \sup_{R(\rho,\delta)} |\cdot|$. Let

$$\omega_*(\eta) = \max\{\omega(2\delta_1, \eta|\{\nabla u^r, \nabla u^l\}), \lambda_*(\eta), \eta\},$$

where $\lambda_*(\eta) = \omega(\varepsilon, \eta|\nabla_u \lambda)$. Now define

$$\begin{aligned} M(\delta) = \{\rho \in M_0(\delta) : \omega(\tau_1(\rho,\delta), \eta|\rho_1') \le K_1\omega_*(\eta), \\ \omega(\tau_2(\rho,\delta), \eta|\rho_2') \le K_2\omega_*(\eta)\}, \end{aligned}$$

where K_1, K_2 are constants to be chosen later. It is not hard to show that $M(\delta)$ is complete in the supremum norm. Define

$$\tilde{\delta}(\rho,\delta) = \sup\{\delta' : R(\tilde{\rho},\delta') \subseteq R(\rho,\delta)\}$$

(see Fig. 4.2).

LEMMA 4.1. *There are constants K_1, K_2, and $\delta_* \in (0, \delta_1]$ such that $S(\rho) \in M(\tilde{\delta}(\rho,\delta))$ for all $\rho \in M(\delta), \delta \le \delta_*$.*

Proof. Differentiating (4.2) we have

$$\tilde{\rho}_1'(y) = -\frac{\partial_y F(y, \tilde{\rho}_1(y))}{\partial_z F(y, \tilde{\rho}_1(y))}, \qquad \tilde{\rho}_2'(z) = -\frac{\partial_z G(\tilde{\rho}_2(z), z)}{\partial_y G(\tilde{\rho}_2(z), z)}. \tag{4.4}$$

Setting $y, z = 0$ and recalling Lemma 1.1, we have that $\tilde{\rho}'(0) = (\rho_1^0, \rho_2^0)$. From the proof of Theorem 3.1, we have

$$\Omega(\delta_1, \eta|\nabla u(\rho)) \le K_3\omega_*(\eta) \quad \text{for all } \rho \in M(\delta_1), \tag{4.5}$$

where K_3 depends on $(1 - |\lambda|_\varepsilon)^{-1}, \Omega_1, |\nabla_u \lambda|_\varepsilon, K_1, K_2$. By (4.4), (4.5), and properties of the Ω, we can choose $\delta_2 \in (0, \delta_1]$ and $\mu < 1$ such that

$$\kappa < \tilde{\rho}_1', \tilde{\rho}_2' < \mu \quad \text{for all } \rho \in M(\delta_2). \tag{4.6}$$

Applying properties of the Ω to (4.4) we have

$$\begin{aligned} \omega(\tilde{\tau}_1, \eta|\tilde{\rho}_1') &\le \|\partial_y F\|_\delta \|1/\partial_z F\|_\delta^2 \omega(\tilde{\tau}_1, \eta|\partial_z F(y, \tilde{\rho}_1(y))) \\ &\quad + \|1/\partial_z F\|_\delta \omega(\tilde{\tau}_1, \eta|\partial_y F(y, \tilde{\rho}_1(y))), \\ \omega(\tilde{\tau}_2, \eta|\tilde{\rho}_2') &\le \|\partial_z G\|_\delta \|1/\partial_y G\|_\delta^2 \omega(\tilde{\tau}_2, \eta|\partial_y G(\tilde{\rho}_2(z), z)) \\ &\quad + \|1/\partial_y G\|_\delta \omega(\tilde{\tau}_2, \eta|\partial_z G(\tilde{\rho}_2(z), z)), \end{aligned} \tag{4.7}$$

where $\tilde{\tau}_i = \tau_i(\tilde{\rho}, \tilde{\delta}(\rho,\delta)), i = 1, 2$. Notice that $u(\rho)$ satisfies (3.20) with $\tilde{\lambda} = \lambda(u(\rho)), \psi = (u_1^r, u_2^l)$. Estimating as in Lemma 3.2 (but more carefully), we can show that

$$\begin{aligned} &\omega(\tilde{\tau}_1, \eta|\partial_y F(y, \tilde{\rho}_1(y))), \omega(\tilde{\tau}_2, \eta|\partial_z G(\tilde{\rho}_2(z), z)) \le (K_4 + K_5\delta)\omega_*(\eta), \\ &\omega(\tilde{\tau}_2, \eta|\partial_y G(\tilde{\rho}_2(z), z)) \le \|\partial_z u_1^r\|_\delta \omega(\tilde{\tau}_2, \eta|\rho_1'(\xi_*(\tilde{\rho}_2(z), z))) \\ &\qquad\qquad + (K_4 + K_5\delta)\omega_*(\eta), \\ &\omega(\tilde{\tau}_1, \eta|\partial_z F(y, \tilde{\rho}_1(y))) \le \|\partial_y u_2^l\|_\delta \omega(\tilde{\tau}_1, \eta|\rho_2'(\zeta_*(y, \tilde{\rho}_1(y)))) \\ &\qquad\qquad + (K_4 + K_5\delta)\omega_*(\eta) \end{aligned} \tag{4.8}$$

for all $\rho \in M(\delta), \delta \le \delta_2$, where ξ_*, ζ_* are as in §3 with $\tilde{\lambda} = \lambda(u(\rho))$. K_4, K_5 depend on $\|\nabla u^r\|_{\delta_1}, \|\nabla u^l\|_{\delta_1}, (1 - |\lambda|_\varepsilon)^{-1}, \Omega_1, |\nabla_u \lambda|_\varepsilon$; K_5 also depends on K_1, K_2.

We can use (3.21), (3.22), (4.6) to show that there is some $\delta_3 \in (0, \delta_2]$ such that

$$\begin{aligned} \left|\frac{d}{dy}\zeta_*(y, \tilde{\rho}_1(y))\right|_{\tilde{\tau}_1} &\le \|\partial_y \zeta_*\|_\delta + |\tilde{\rho}_1'|_{\tilde{\tau}_1}\|\partial_z \zeta_*\|_\delta \le 1, \\ \left|\frac{d}{dz}\xi_*(\tilde{\rho}_2(z), z)\right|_{\tilde{\tau}_2} &\le \|\partial_z \xi_*\|_\delta + |\tilde{\rho}_2'|_{\tilde{\tau}_2}\|\partial_y \xi_*\|_\delta \le 1 \end{aligned} \tag{4.9}$$

for all $\rho \in M(\delta), \delta \le \delta_3$. Combining (4.7), (4.8), and (4.9) and using properties of ω, we have

$$\begin{aligned} \omega(\tilde{\tau}_1, \eta|\tilde{\rho}_1') &\le B_1(\delta)K_2\omega_*(\eta) + C_1(\delta)\omega_*(\eta), \\ \omega(\tilde{\tau}_2, \eta|\tilde{\rho}_2') &\le B_2(\delta)K_1\omega_*(\eta) + C_2(\delta)\omega_*(\eta) \end{aligned} \tag{4.10}$$

for all $\rho \in M(\delta), \delta \le \delta_3$, where

$$\begin{aligned} B_1(\delta) &= \|\partial_y F\|_\delta \|1/\partial_z F\|_\delta^2 \|\partial_y u_2^l\|_\delta, \\ B_2(\delta) &= \|\partial_z G\|_\delta \|1/\partial_y G\|_\delta^2 \|\partial_z u_1^r\|_\delta, \\ C_1(\delta) &= (\|\partial_y F\|_\delta \|1/\partial_z F\|_\delta^2 + \|1/\partial_z F\|_\delta)(K_4 + K_5\delta), \\ C_2(\delta) &= (\|\partial_z G\|_\delta \|1/\partial_y G\|_\delta^2 + \|1/\partial_y G\|_\delta)(K_4 + K_5\delta). \end{aligned}$$

Notice that $C_1(0), C_2(0)$ do not depend on K_1, K_2, so we can choose $C > C_1(0), C_2(0)$ and let

$$K_1 = \frac{(B_1(0)+1)C}{1 - B_1(0)B_2(0)}, \qquad K_2 = \frac{(B_2(0)+1)C}{1 - B_1(0)B_2(0)}. \tag{4.11}$$

(Using (2.2), (2.6), (3.1), (3.2), (3.4), and (3.35), one can show that $B_1(0)B_2(0) = (\rho_1^0\rho_2^0)^2(1-c_0)^2/(1+c_0)^2 < 1$.)

We need to show from (4.10) that for δ sufficiently small

$$\omega(\tilde{\tau}_1, \eta|\tilde{\rho}_1') \le K_1\omega_*(\eta), \qquad \omega(\tilde{\tau}_2, \eta|\tilde{\rho}_2') \le K_2\omega_*(\eta), \tag{4.12}$$

which can be accomplished by showing that

$$B_1(\delta)K_2 + C_1(\delta) \le K_1, \qquad B_2(\delta)K_1 + C_2(\delta) \le K_2.$$

Using (4.11) we can show that this is equivalent to

$$\begin{aligned} (B_1(\delta) - B_1(0))(B_2(0)+1)C &\le (C - C_1(\delta))(1 - B_1(0)B_2(0)), \\ (B_2(\delta) - B_2(0))(B_1(0)+1)C &\le (C - C_2(\delta))(1 - B_1(0)B_2(0)). \end{aligned} \tag{4.13}$$

By (4.5) there is some $\delta_* \in (0, \delta_3]$ such that (4.13), and therefore (4.12), hold for all $\rho \in M(\delta), \delta \le \delta_*$. This completes the proof. □

It is clear from the proof of Lemma 4.1 that the curves of $\tilde{\rho}$ intersect the "outer" edges of $R(\rho, \delta)$, i.e., those that meet the curves of ρ, so $\tilde{\delta}(\rho, \delta)$ is determined by where these intersections occur. Figure 4.2 shows an example of this.

It is not hard to show that

$$\delta - \tilde{\delta}(\rho, \delta) \le \frac{2\kappa}{1 - \kappa^2}|\tilde{\rho} - \rho|_{\tilde{\tau}} \tag{4.14}$$

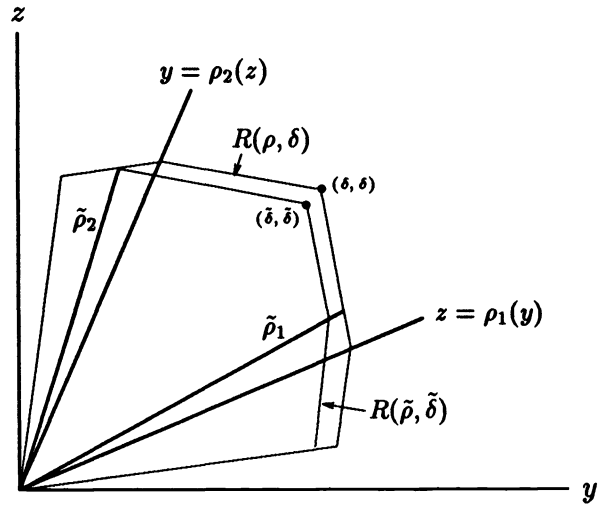


FIG. 4.2.

for all $\rho \in M(\delta), \delta \le \delta_*$, where $|\rho|_\tau = \max_{i=1,2} |\rho_i|_{\tau_i}$.

Let $\rho^{(0)} \in M(\delta_0), \delta_0 \le \delta_*$, and consider the sequence

$$\rho^{(n)} = S(\rho^{(n-1)}) \in M(\delta_n), \qquad n = 1, 2, \ldots,$$

where $\delta_n = \tilde{\delta}(\rho^{(n-1)}, \delta_{n-1})$, and $\rho_i^{(n)}$ is restricted to $[0, \tau_i(\rho^{(n)}, \delta_n)], i = 1, 2$.

THEOREM 4.1. *There is some $\delta_L \in (0, \delta_*]$ such that $\rho^{(n)} \to \rho^L \in M(\delta_L)$ as $n \to \infty$ and ρ^L is the unique fixed point of S satisfying $\tilde{\rho}'(0) = (\rho_1^0, \rho_2^0)$.*

We need the following lemma.

LEMMA 4.2. *There is some $\delta^* \in (0, \delta_*]$ and $\nu \in (0, 1)$ such that*

$$|\rho^{(n+3)} - \rho^{(n+2)}|_{\tau^{(n+3)}} \le \nu |\rho^{(n+1)} - \rho^{(n)}|_{\tau^{(n+1)}}$$

for all $\rho^{(0)} \in M(\delta_0), \delta_0 \le \delta^*$, where $\tau_i^{(n)} = \tau_i(\rho^{(n)}, \delta_n), i = 1, 2$.

Proof. Let $F^{(n)} = u_2(\rho^{(n)}) - u_2^r, G^{(n)} = u_1(\rho^{(n)}) - u_1^l, u^* = u(\rho^{(n+1)}) - u(\rho^{(n)})$. By (4.2) we have

$$F^{(n)}(y, \rho_1^{(n+2)}(y)) - F^{(n)}(y, \rho_1^{(n+1)}(y)) + u_2^*(y, \rho_1^{(n+2)}(y)) = 0,$$
$$G^{(n)}(\rho_2^{(n+2)}(z), z) - G^{(n)}(\rho_2^{(n+1)}(z), z) + u_1^*(\rho_2^{(n+2)}(z), z) = 0,$$

which implies that

$$\begin{aligned} \rho_1^{(n+2)}(y) - \rho_1^{(n+1)}(y) &= -u_2^*(y, \rho_1^{(n+2)}(y))/\partial_z F^{(n)}(y, \zeta), \\ \rho_2^{(n+2)}(z) - \rho_2^{(n+1)}(z) &= -u_1^*(\rho_2^{(n+2)}(z), z)/\partial_y G^{(n)}(\xi, z) \end{aligned} \tag{4.15}$$

where ζ is between $\rho_1^{(n+1)}(y), \rho_1^{(n+2)}(y)$ and ξ is between $\rho_2^{(n+1)}(z), \rho_2^{(n+2)}(z)$. By (3.5), (4.1), u^* satisfies (3.31) on $R(\rho^{(n+1)}, \delta_{n+1})$ with $u^*, u(\rho^{(n)}), u(\rho^{(n+1)})$ in place of $\tilde{u}^*, u^{(1)}, u^{(2)}$, and

$$\begin{aligned} u_1^* &= u_1^r - u_1(\rho^{(n)}) \quad \text{on } z = \rho_1^{(n+1)}(y), \\ u_2^* &= u_2^l - u_2(\rho^{(n)}) \quad \text{on } y = \rho_2^{(n+1)}(z). \end{aligned} \tag{4.16}$$

Using (4.1) we can rewrite (4.16) as

$$u_1^*(y,\rho_1^{(n+1)}(y))=h_1(y),\qquad u_2^*(\rho_2^{(n+1)}(z),z)=h_2(z), \tag{4.17}$$

where

$$\begin{aligned} h_1(y)&=u_1(\rho^{(n)})(y,\rho_1^{(n)}(y))-u_1(\rho^{(n)})(y,\rho_1^{(n+1)}(y))\\ &\quad+u_1^r(y,\rho_1^{(n+1)}(y))-u_1^r(y,\rho_1^{(n)}(y)),\\ h_2(z)&=u_2(\rho^{(n)})(\rho_2^{(n)}(z),z)-u_2(\rho^{(n)})(\rho_2^{(n+1)}(z),z)\\ &\quad+u_2^l(\rho_2^{(n+1)}(z),z)-u_2^l(\rho_2^{(n)}(z),z). \end{aligned} \tag{4.18}$$

By rewriting (3.31), (4.17) in an integral form and using (4.3), we can show that

$$\begin{aligned} \|u_1^*\|_{\delta_{n+1}}&\le(1+A_0\delta_n)|h_1|_{\tau_1^{(n+1)}}\\ &\quad+A_0\delta_n(|h_2|_{\tau_2^{(n+1)}}+\|u^*\|_{\delta_{n+1}}+\|H\|_{\delta_{n+1}}),\\ \|u_2^*\|_{\delta_{n+1}}&\le(1+A_0\delta_n)|h_2|_{\tau_2^{(n+1)}}\\ &\quad+A_0\delta_n(|h_1|_{\tau_1^{(n+1)}}+\|u^*\|_{\delta_{n+1}}+\|H\|_{\delta_{n+1}}) \end{aligned} \tag{4.19}$$

for $\delta_0\le\delta_*$, where $\|\cdot\|_{\delta_n}=\sup_{R(\rho^{(n)},\delta_n)}|\cdot|$ and A_0 depends on $\Omega_1,(1-|\lambda|_\varepsilon)^{-1},|\nabla_u\lambda|_\varepsilon$. By (3.34) we have

$$\|H\|_{\delta_{n+1}}\le A_1\Omega_1\|u^*\|_{\delta_{n+1}}, \tag{4.20}$$

where A_1 depends on $|\nabla_u\lambda|_\varepsilon$. From (4.18) we can see that

$$\begin{aligned} |h_1|_{\tau_1^{(n+1)}}&\le\|\partial_z u_1(\rho^{(n)})-\partial_z u_1^r\|_{\delta_n}|\rho_1^{(n+1)}-\rho_1^{(n)}|_{\tau_1^{(n+1)}},\\ |h_2|_{\tau_2^{(n+1)}}&\le\|\partial_y u_2(\rho^{(n)})-\partial_y u_2^l\|_{\delta_n}|\rho_2^{(n+1)}-\rho_2^{(n)}|_{\tau_2^{(n+1)}} \end{aligned} \tag{4.21}$$

for $\delta_0\le\delta_*$. Equations (4.19)–(4.21) imply that there is some $\delta'\in(0,\delta_*]$ such that for $\delta_0\le\delta'$

$$\begin{aligned} \|u_1^*\|_{\delta_{n+1}}&\le(1+A_2\delta_n)\|\partial_z u_1(\rho^{(n)})-\partial_z u_1^r\|_{\delta_n}|\rho_1^{(n+1)}-\rho_1^{(n)}|_{\tau_1^{(n+1)}}\\ &\quad+A_2\delta_n\|\partial_y u_2(\rho^{(n)})-\partial_y u_2^l\|_{\delta_n}|\rho_2^{(n+1)}-\rho_2^{(n)}|_{\tau_2^{(n+1)}},\\ \|u_2^*\|_{\delta_{n+1}}&\le(1+A_2\delta_n)\|\partial_y u_2(\rho^{(n)})-\partial_y u_2^l\|_{\delta_n}|\rho_2^{(n+1)}-\rho_2^{(n)}|_{\tau_2^{(n+1)}}\\ &\quad+A_2\delta_n\|\partial_z u_1(\rho^{(n)})-\partial_z u_1^r\|_{\delta_n}|\rho_1^{(n+1)}-\rho_1^{(n)}|_{\tau_1^{(n+1)}} \end{aligned} \tag{4.22}$$

for all $\rho^{(0)}\in M(\delta_0)$, where A_2 depends on Ω_1,A_0,A_1. Combining this with (4.15), we have

$$\begin{aligned} |\rho_1^{(n+2)}-\rho_1^{(n+1)}|_{\tau_1^{(n+2)}}&\le B_1^{(n)}|\rho_2^{(n+1)}-\rho_2^{(n)}|_{\tau_2^{(n+1)}}+C_1^{(n)}\delta_n|\rho_1^{(n+1)}-\rho_1^{(n)}|_{\tau_1^{(n+1)}},\\ |\rho_2^{(n+2)}-\rho_2^{(n+1)}|_{\tau_2^{(n+2)}}&\le B_2^{(n)}|\rho_1^{(n+1)}-\rho_1^{(n)}|_{\tau_1^{(n+1)}}+C_2^{(n)}\delta_n|\rho_2^{(n+1)}-\rho_2^{(n)}|_{\tau_2^{(n+1)}} \end{aligned} \tag{4.23}$$

for $\delta_0 \le \delta'$, where

$$B_1^{(n)} = \|1/\partial_z F^{(n)}\|_{\delta_n} \|\partial_y u_2(\rho^{(n)}) - \partial_y u_2^l\|_{\delta_n}(1 + A_2\delta_n),$$
$$B_2^{(n)} = \|1/\partial_y G^{(n)}\|_{\delta_n} \|\partial_z u_1(\rho^{(n)}) - \partial_z u_1^r\|_{\delta_n}(1 + A_2\delta_n),$$
$$C_1^{(n)} = \|1/\partial_z F^{(n)}\|_{\delta_n} \|\partial_z u_1(\rho^{(n)}) - \partial_z u_1^r\|_{\delta_n} A_2,$$
$$C_2^{(n)} = \|1/\partial_y G^{(n)}\|_{\delta_n} \|\partial_y u_2(\rho^{(n)}) - \partial_y u_2^l\|_{\delta_n} A_2.$$

This implies that

$$|\rho^{(n+3)} - \rho^{(n+2)}|_{\tau^{(n+3)}} \le (B_1B_2 + B_1C_2\delta_0 + B_2C_1\delta_0 + B_1C_1\delta_0 + B_2C_2\delta_0 + C_1^2\delta_0^2)|\rho^{(n+1)} - \rho^{(n)}|_{\tau^{(n+1)}}$$

for $\delta_0 \le \delta'$, where $B_i = \sup_{n\ge 0} B_i^{(n)}, C_i = \sup_{n\ge 0} C_i^{(n)}, i = 1, 2$. Using (2.2), (2.6), (3.1), (3.2), (3.4), (3.35), and (4.5), one can show that

$$\lim_{\delta_0 \to 0} B_1B_2 = \rho_1^0\rho_2^0\left(\frac{1-c_0}{1+c_0}\right)^2 < 1.$$

Choose ν such that

$$\rho_1^0\rho_2^0\left(\frac{1-c_0}{1+c_0}\right)^2 < \nu < 1.$$

Because of (4.5), there is some $\delta^* \in (0, \delta']$ such that

$$(B_1B_2 + B_1C_2\delta_0 + B_2C_1\delta_0 + B_1C_1\delta_0 + B_2C_2\delta_0 + C_1^2\delta_0^2) \le \nu$$

for all $\rho^{(0)} \in M(\delta_0), \delta_0 \le \delta^*$. This completes the proof. □

Proof of Theorem 4.1. Notice that because of (4.14)

$$\delta_0 - \delta_n \le \sum_{k=0}^{\infty} \frac{2\kappa}{1-\kappa^2}|\rho^{(k+1)} - \rho^{(k)}|_{\tau^{(k+1)}}, \qquad n \ge 0,$$

for $\delta_0 \le \delta^*$. So by Lemma 4.2

$$\begin{aligned}\delta_0 - \delta_n &\le (|\rho^{(2)} - \rho^{(1)}|_{\tau^{(2)}} + |\rho^{(1)} - \rho^{(0)}|_{\tau^{(1)}})2\kappa/(1-\nu)(1-\kappa^2) \\ &\le 4\kappa\delta_0/(1-\nu)(1-\kappa^2).\end{aligned}$$

Assume κ, ν were chosen so that $4\kappa/(1-\nu)(1-\kappa^2) < 1$. Then there is some $\delta_L \in (0, \delta_0]$ such that $\rho^{(n)} \in M(\delta_L), n \ge 0$. Now Lemma 4.2 and the completeness of $M(\delta_L)$ imply that $\rho^{(n)} \to \rho^L \in M(\delta_L)$ as $n \to \infty$ and $S(\rho^L) = \rho^L$. To prove uniqueness, suppose we have another solution ρ. Let $\bar{\delta} \le \delta_L$ be maximal such that $\kappa \le \rho_1', \rho_2' \le 1$ on $R(\rho, \bar{\delta})$. We can slightly modify the proof of Lemma 4.2 to show that $\rho^L = \rho$ on $R(\rho, \bar{\delta})$. We know that $\bar{\delta} > 0$ since $\rho'(0) = (\rho_1^0, \rho_2^0)$, so $\kappa < \rho_1', \rho_2' < 1$ on $R(\rho, \bar{\delta})$ by (4.6). This implies that $\bar{\delta} = \delta_L$, completing the proof. □

Remark. We now modify the methods of this section to complete the argument for Case vi in §2. Changing variables as in (3.1), (3.2), the free boundary problem in Case vi is equivalent to (3.5) with boundary conditions

$$u = u^r \quad \text{on } z = \rho(y), \qquad u = u^l \quad \text{on } y = \mu(z)$$

where ρ is unknown, u^r is as in (3.4), u^l is the transformed solution on P_1, and $y = \mu(z)$ is the transformed characteristic boundary of P_2, so $\mu' = \lambda(u^l)$ on $y = \mu(z)$.

Let $\rho \in C^1[0,\delta], \rho'(0) = \rho_0$, where $z = \rho_0 y$ is the right-moving interface in the scale-invariant solution. Define $u(\rho)$ to be the solution to (3.5) satisfying boundary conditions

$$u_1(\rho) = u_1^r \quad \text{on } z = \rho(y), \qquad u(\rho) = u^l \quad \text{on } y = \mu(z).$$

The local existence of $u(\rho)$ follows from Theorem 4.3 of [3], which is proven in essentially the same manner as Theorem 3.1. The only difference is that, since u_2^l is used as the left boundary condition, it must be verified that the solution obtained satisfies $u_1 = u_1^l$ on $y = \mu(z)$. Define $S(\rho) = \tilde{\rho}$ where $u_2^r(y, \tilde{\rho}(y)) = u_2(\rho)(y, \tilde{\rho}(y))$. As in Case ii, we seek a fixed point ρ^L of S. We can modify the argument in this section as follows.

Choose $\kappa \in (0, \rho_0)$, define $\tau_1(\rho, \delta)$ as before with ρ, μ in place of ρ_1, ρ_2, and define $R(\rho, \delta)$ by

$$R(\rho,\delta) = \{(y,z) : \mu(z) \le y \le \pm\kappa(z - \rho(\tau_1(\delta))) + \tau_1(\delta),$$
$$\kappa y \le z \le -\kappa(y - \mu(\tau_2(\delta))) + \tau_2(\delta)\}.$$

(This is analogous to the previous definition of $R(\rho, \delta)$ leaving out the portion to the left of $y = \mu(z)$.) Define

$$M_0(\delta) = \{\rho \in C^1[0, \tau_1(\rho,\delta)]: \rho(0) = 0, \rho'(0) = \rho_0, \kappa \le \rho' \le 1\},$$
$$\omega_*(\eta) = \max\{\omega(2\delta_1, \eta | \{\nabla u^r, \nabla u^l, \mu'\}), \lambda_*(\eta), \eta\}.$$

Omit all references to ρ_2, K_2, and G. Replace ρ_1 by ρ and τ, τ_i by τ_1 throughout. In place of (4.8), we have

$$\omega(\tilde{\tau}_1, \eta | \nabla F(y, \tilde{\rho}(y))) \le (K_4 + K_5\delta)\omega_*(\eta),$$

where K_4, K_5 are as in the proof of Lemma 4.1. Hence, in place of (4.10), we have that

$$\omega(\tilde{\tau}_1, \eta | \tilde{\rho}') \le C_1(\delta)\omega_*(\eta),$$

where C_1 is as in the proof of Lemma 4.1. So, to prove Lemma 4.1 in this case, we just choose $K_1 > C_1(0)$ and δ_* such that $C_1(\delta_*) \le K_1$ for all $\rho \in M(\delta_*)$.

In (4.16), replace $\rho_2^{(n+1)}$ by μ and notice that $h_2 \equiv 0$. Hence, in place of (4.22), we have

$$\|u_2^*\|_{\delta_{n+1}} \le A_2\delta_n \|\partial_z u_1(\rho^{(n)}) - \partial_z u_1^r\|_{\delta_n} |\rho^{(n+1)} - \rho^{(n)}|_{\tau_1^{(n+1)}},$$

and so in place of (4.23) we have

$$|\rho^{(n+2)} - \rho^{(n+1)}|_{\tau_1^{(n+2)}} \le C_1\delta_n |\rho^{(n+1)} - \rho^{(n)}|_{\tau_1^{(n+1)}},$$

where A_2, C_1 are as in the proof of Lemma 4.2. Now choose δ^* such that $C_1\delta^* < \nu$ for all $\rho^{(0)} \in M(\delta^*)$. The rest of the argument is the same.

Acknowledgments. I would like to thank my advisor, Dr. David G. Schaeffer, for his support and supervision of this research, and Dr. Michael Shearer for suggesting the problem.

REFERENCES

[1] S. ANTMAN AND W. SZYMCZAK, *Nonlinear elastoplastic waves*, Contemp. Math., 100 (1989), pp. 27–54.

[2] R. CLIFTON AND S. BODNER, *An analysis of elastic-plastic pulse propagation*, J. Appl. Mech., 33 (1966), pp. 248–255.

[3] MICHAEL K. GORDON, *Perturbed Scale-Invariant Initial Value Problems in One-Dimensional Dynamic Elastoplasticity*, Ph.D. thesis, Duke University, Durham, NC, 1993.

[4] E. LEE, *A boundary value problem in the theory of plastic wave propagation*, Quart. Appl. Math., 10 (1952), pp. 335–346.

[5] D. SCHAEFFER AND M. SHEARER, *Scale-invariant initial value problems in one-dimensional dynamic elastoplasticity, with consequences for multidimensional nonassociative plasticity*, European J. Appl. Math., 3 (1992), pp. 225–254.

[6] LI TA-TSIEN AND YU WEN-CI, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke University Mathematics Series 5, Duke University Mathematics Department, Durham, NC, 1985.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1588–1610, November 1995

012

COUPLED PARABOLIC AND HYPERBOLIC EQUATIONS MODELING AGE-DEPENDENT EPIDEMIC DYNAMICS WITH NONLINEAR DIFFUSION*

CHAOCHENG HUANG† AND JIONGMIN YONG‡

Dedicated to Professor Avner Friedman at his 60th birthday

Abstract. This paper considers a system of coupled second-order parabolic and first-order hyperbolic equations arising from the age-dependent diffusion population dynamics with an infectious disease. The diffusion is assumed to be nonlinear, which leads to the parabolic equation being degenerate. A notion of weak solutions is introduced. Under mild conditions, the authors have proved the global existence of weak solutions. The result is further improved for the one-dimensional case.

Key words. population model, nonlinear diffusion, coupled parabolic, hyperbolic equations

AMS subject classifications. 35D05, 35K65, 35L60, 92D25, 92D30

1. Introduction.

In this paper we study the population dynamics of a single species with an infectious disease. The population is divided into two groups: susceptibles (who can catch the disease) and infectives (who can infect the disease). We consider the problem in the whole space $\mathbb{R}^n$ (in practice, $n \leq 3$) and whole time interval $[0, \infty)$. By $(x,t) \in \mathbb{R}^n \times [0,\infty)$ we mean "at location x and at time t." We let $a \in [0,\infty)$ be the age of the infectives since catching the disease. Next, we let $\rho(x,t,a)$ be the age distribution of the infectives. Roughly speaking, this is the number of infectives who have caught the disease for time length a (time units, say days or hours) at location x and at time t. Thus, the density of the infectives at $(x,t) \in \mathbb{R}^n \times [0,\infty)$ is

$$u(x,t) \equiv \int_0^\infty \rho(x,t,a)da. \tag{1.1}$$

The density of the susceptibles at (x,t) is denoted by $v(x,t)$. In this paper we do not consider the birth and natural death. Such a mixed situation will be considered in our future works. We let $\lambda(x,t,a)$, $\beta(x,t,a)$, and $\gamma(x,t,a)$ be the death rate for the infectives from the disease, the recovery rate, and the infection rate at (x,t,a), respectively. If there is no diffusion with respect to the space variable $x \in \mathbb{R}^n$, then all the functions are independent of x and we have the following equations for ρ and

*Received by the editors October 19, 1993; accepted for publication (in revised form) February 18, 1994. This work was completed while the second author was visiting the Institute for Mathematics and Its Applications (IMA), University of Minnesota, Minneapolis, Minnesota. This research was partially supported by the Institute for Mathematics and Its Applications, National Science Foundation of China grant 19131050, and the Fok Ying Tung Education Foundation.

†School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

‡Department of Mathematics, Fudan University, Shanghai 200433, China.

v ($\mathbb{R}^+ = (0, \infty)$):

$$
\begin{cases}
\rho_t + \rho_a = -\lambda(t,a)\rho - \beta(t,a)\rho, & (t,a) \in \mathbb{R}^+ \times \mathbb{R}^+, \\
v_t = -\left(\int_0^\infty \gamma(t,a)\rho(t,a)da\right) v + \int_0^\infty \beta(t,a)\rho(t,a)da, & t \in \mathbb{R}^+, \\
\rho\,|_{t=0} = \rho_0(a), & a \in \mathbb{R}^+, \\
\rho\,|_{a=0} = \left(\int_0^\infty \gamma(t,a)\rho(t,a)da\right) v(t), & t \in \mathbb{R}^+, \\
v\,|_{t=0} = v_0.
\end{cases} \tag{1.2}
$$

See [15], [2], [11], and [17] for the relevant details. Now we are interested in the case in which the population is also subject to space diffusion. We assume that the diffusion is a result of overcrowding. From the model introduced in [9] (also see [1], [3], [4], [8], [10], [13]–[15]), the diffusion velocity can be taken as $-\nabla(u+v)$. Here we notice that $u+v$ is the density of the total population. It is clear that to take this diffusion effect into account, we need to add terms $\nabla \cdot [\rho\nabla(u+v)]$ and $\nabla \cdot [v\nabla(u+v)]$ into the first two equations in (1.2), respectively. This can be justified by the conservation of the population in both of the two groups (infectives and susceptibles). Hence, we end up with the following system:

$$
\begin{aligned}
\rho_t + \rho_a = \nabla \cdot [\rho\nabla(u+v)] - \lambda(x,t,a)\rho - \beta(x,t,a)\rho, \\
(x,t,a) \in \mathbb{R}^n \times \mathbb{R}^+ \times \mathbb{R}^+,
\end{aligned} \tag{1.3}
$$

(1.4)

$$
\begin{aligned}
v_t = \nabla \cdot [v\nabla(u+v)] - \left(\int_0^\infty \gamma(x,t,a)\rho(x,t,a)da\right) v + \int_0^\infty \beta(x,t,a)\rho(x,t,a)da, \\
(x,t) \in \mathbb{R}^n \times \mathbb{R}^+
\end{aligned}
$$

with the following initial conditions:

$$
\rho\,|_{t=0} = \rho_0(x,a), \qquad (x,a) \in \mathbb{R}^n \times \mathbb{R}^+, \tag{1.5}
$$

$$
\rho\,|_{a=0} = \left(\int_0^\infty \gamma(x,t,a)\rho(x,t,a)da\right) v(x,t), \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}^+, \tag{1.6}
$$

and

$$
v\,|_{t=0} = v_0(x), \qquad x \in \mathbb{R}^n. \tag{1.7}
$$

Now let us further simplify the model to catch the essence of it. To this end we let λ, β, and γ be independent of a. Then the density u of the infectives defined in (1.1) and the density v of the susceptibles satisfy the following coupled system:

$$
u_t = \nabla \cdot [u\nabla(u+v)] + \gamma(x,t)uv - [\lambda(x,t) + \beta(x,t)]u, \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}^+, \tag{1.8}
$$

$$
v_t = \nabla \cdot [v\nabla(u+v)] - \gamma(x,t)uv + \beta(x,t)u, \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}^+. \tag{1.9}
$$

If we set

$$p = \begin{pmatrix} u \\ v \end{pmatrix}, \qquad A = \begin{pmatrix} u & u \\ v & v \end{pmatrix},$$

equations (1.8)–(1.9) can be written as

$$p_t - A\Delta p = G(x, t, p, \nabla p) \tag{1.10}$$

with some nonlinear function G. This system is not parabolic since the matrix $(A + A^T)/2$ has a negative eigenvalue if $u \neq v$. To overcome this difficulty, we introduce a new variable $w = u + v$. This is nothing but the density of the total population. Then system (1.8)–(1.9) can be transformed into the following form:

$$w_t = \nabla \cdot (w \nabla w) - \lambda u, \qquad (x, t) \in \mathbb{R}^n \times \mathbb{R}^+, \tag{1.11}$$

$$u_t = \nabla u \cdot \nabla w + u\Delta w + (\gamma w - \lambda - \beta)u - \gamma u^2, \qquad (x, t) \in \mathbb{R}^n \times \mathbb{R}^+ \tag{1.12}$$

with initial conditions

$$u\,|_{t=0} = u_0(x) \equiv \int_0^\infty \rho_0(x, a)da, \qquad x \in \mathbb{R}^n, \tag{1.13}$$

$$w\,|_{t=0} = w_0(x) \equiv u_0(x) + v_0(x), \qquad x \in \mathbb{R}^n. \tag{1.14}$$

For fixed u, (1.11) is an inhomogeneous porous medium equation and for fixed w, (1.12) is a first-order nonlinear hyperbolic equation, which is not of conservation law type. Hence (1.11)–(1.14) is a Cauchy problem for a system of a second-order degenerate parabolic equation coupled with a first-order hyperbolic equation. For the porous medium equation we know that, in general, there exist no classical solutions if the initial data is not strictly positive everywhere. Thus we cannot expect to have a classical solution for (1.11)–(1.14). In this paper, we study the weak solutions for this system. In §2 we introduce the weak formulation of (1.11)–(1.14). Section 3 is devoted to the study of the approximate problems. The existence of weak solutions will be proved in §4. In §5, the age distribution function ρ for the infectives is recovered from (1.3). Finally, in §6 the one-dimensional case is discussed.

2. A weak formulation. In this section we introduce a weak formulation of (1.11)–(1.14). Let us first introduce some notation. Let $\Omega = \mathbb{R}^n$. For any $T > 0$ we denote $\Omega_T = (0, T) \times \mathbb{R}^n$, $\bar{\Omega}_T^0 = [0, T) \times \mathbb{R}^n$, and $\bar{\Omega}_T = [0, T] \times \mathbb{R}^n$. For any integer $k \geq 0$, let $C^k(\Omega)$ be the set of all k-time continuously differentiable functions $u(x)$ defined on Ω with all partial derivatives up to order k (inclusively) being bounded. The norm of $C^k(\Omega)$ is denoted by $\|\cdot\|_{C^k(\Omega)}$. For $k \geq 0$ and $0 < \alpha < 1$, we denote $C^{k+\alpha}(\Omega)$ as the Banach space of all functions $u(x)$ in $C^k(\Omega)$ with the kth-order partial derivatives being α-Hölder continuous. The norm of this space is denoted by $\|\cdot\|_{C^{k+\alpha}(\Omega)}$. We can define the spaces $C^k(\bar{\Omega}_T)$ and $C^{k+\alpha}(\bar{\Omega}_T)$ in a similar way, treating x and t equally. Finally, we let $C_0^\infty(\bar{\Omega}_T^0)$ be the set of all functions ξ which belong to $C^k(\bar{\Omega}_T)$ for any $k \geq 0$, and satisfy $\xi(T, x) = 0$ for all $x \in \mathbb{R}^n$ and $\xi(t, x) = 0$ for $t \geq 0$ and $|x| \geq M$ (for some $M > 0$).

Next we need to introduce spaces which are suitable for parabolic problems. To this end, we define the parabolic distance as follows:

$$d((x, t), (x', t')) = (|x - x'|^2 + |t - t'|)^{1/2} \quad \forall (x, t), (x', t') \in \mathbb{R}^n \times \mathbb{R}^+. \tag{2.1}$$

For $0 < \alpha \leq 1$, we denote $C^{0,\alpha}(\bar{\Omega}_T)$ as the space of all functions $u(x,t)$ in $C^0(\bar{\Omega}_T)$ for which

$$\|u\|_{C^{0,\alpha}(\bar{\Omega}_T)} \equiv \|u\|_{C^0(\bar{\Omega}_T)} + \sup_{(x,t),(x't')\in\bar{\Omega}_T,(x,t)\neq(x't')} \frac{|u(x,t)-u(x',t')|}{d((x,t),(x',t'))^\alpha} \tag{2.2}$$

is finite. We set $C^{0,0}(\bar{\Omega}_T) \equiv C^0(\bar{\Omega}_T)$ with the same norm as $C^0(\bar{\Omega}_T)$. For $0 \leq \alpha \leq 1$, we define the space $C^{1,\alpha}(\bar{\Omega}_T)$ as the set of all $C(\bar{\Omega}_T)$ functions $u(x,t)$ in $C^0(\bar{\Omega}_T)$ having the first-order partial derivatives in x with

$$\|u\|_{C^{1,\alpha}(\bar{\Omega}_T)} \equiv \|u\|_{C^{0,\alpha}(\bar{\Omega}_T)} + \sum_{i=1}^{n} \|u_{x_i}\|_{C^{0,\alpha}(\bar{\Omega}_T)} \tag{2.3}$$

being finite, and we define $C^{2,\alpha}(\bar{\Omega}_T)$ as the set of all functions $u(x,t)$ in $C^1(\bar{\Omega}_T)$ having the second-order partial derivatives in x for which

$$\|u\|_{C^{2,\alpha}(\bar{\Omega}_T)} \equiv \|u\|_{C^{1,\alpha}(\bar{\Omega}_T)} + \|u_t\|_{C^{0,\alpha}(\bar{\Omega}_T)} + \sum_{i,j=1}^{n} \|u_{x_ix_j}\|_{C^{0,\alpha}(\bar{\Omega}_T)} \tag{2.4}$$

is finite.

Throughout of this paper, we make the following hypotheses:

$$\lambda(x,t), \beta(x,t), \gamma(x,t) \in C^1(\mathbb{R}^n \times \mathbb{R}^+), \tag{2.5}$$

$$w_0, u_0 \in C^{2+\alpha}(\mathbb{R}^n), \tag{2.6}$$

$$0 \leq \lambda, \beta, \gamma \leq M \quad \text{for some constant } M, \tag{2.7}$$

$$w_0(x) \geq u_0(x) \geq 0 \quad \forall x \in \mathbb{R}^n. \tag{2.8}$$

Also, either

$$w_0(x) \geq \bar{w}_0 > 0 \quad \forall x \in \mathbb{R}^n \tag{2.9}$$

or $w_0(x)$ has a compact support and satisfies

$$w_0(x) \geq M_0 \text{dist}\,(x, \Gamma_0) \quad \forall x \in \mathbb{R}^n \text{ with } w_0(x) > 0, \tag{2.10}$$

where $M_0 > 0$ is a constant and $\Gamma_0 = \partial\{w_0 > 0\}$ is the boundary of the support of w_0.

The first two assumptions are for convenience, which can be slightly relaxed. The rest of them are assumed for technical reasons; also, they are physically reasonable.

Suppose (w,u) is a classical solution of (1.11)–(1.14). Multiplying (1.11) by any $\xi \in C_0^\infty(\bar{\Omega}_T^0)$ and then integrating by parts, we obtain

$$\int_{\Omega_T} \left\{ w\xi_t - \frac{1}{2}\nabla w^2 \cdot \nabla\xi - \lambda u\xi \right\} dxdt - \int_{\mathbb{R}^n} w_0\xi(0,x)dx = 0. \tag{2.11}$$

On the other hand, (1.12) can be written as

$$u_t = \nabla \cdot [u\nabla w] + (\gamma w - \lambda - \beta)u - \gamma u^2, \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}^+. \tag{2.12}$$

Thus, multiplying (2.12) by any $\eta \in C_0^\infty(\bar{\Omega}_T^0)$ and integrating by parts, we have

$$\int_{\Omega_T} \{u\eta_t - u\nabla w \cdot \nabla \eta + [(\gamma w - \lambda - \beta)u - \gamma u^2]\eta\} dxdt - \int_{\mathbf{R}^n} u_0 \eta(0,x)dx = 0. \tag{2.13}$$

Let us introduce the function $\theta(r,s)$ as follows:

$$\theta(r,s) = \begin{cases} \dfrac{s}{r}, & r > 0, \\ 0, & r \le 0. \end{cases} \tag{2.14}$$

Then we can write (2.13) as follows:

$$\begin{aligned} &\int_{\Omega_T} \left\{ u\eta_t - \frac{1}{2}\theta(w,u)\nabla w^2 \cdot \nabla \eta + [(\gamma w - \lambda - \beta)u - \gamma u^2]\eta \right\} dxdt \\ &\quad - \int_{\mathbf{R}^n} u_0 \eta(0,x)dx = 0. \end{aligned} \tag{2.15}$$

The above analysis suggests that we introduce the following weak formulation for (1.11)–(1.14).

DEFINITION 2.1. *A triple of functions (w,u,χ) defined on Ω_T is called a weak solution of* (1.11)–(1.14) *in Ω_T if the following hold:*

$$w, u, \chi \in L^\infty(\Omega_T), \qquad \nabla w^2 \in L^2_{\text{loc}}(\Omega_T), \tag{2.16}$$

$$w \ge u \ge 0, \qquad w^2 \ge \chi \ge u^2 \ge 0, \tag{2.17}$$

and (w,u,χ) satisfy (2.11) *and*

$$\begin{aligned} &\int_{\Omega_T} \left\{ u\eta_t - \frac{1}{2}\theta(w,u)\nabla w^2 \cdot \nabla \eta + [(\gamma w - \lambda - \beta)u - \gamma \chi]\eta \right\} dxdt \\ &\quad - \int_{\mathbf{R}^n} u_0 \eta(0,x)dx = 0 \end{aligned} \tag{2.18}$$

for all $\xi, \eta \in C_0^\infty(\bar{\Omega}_T^0)$. If (w,u,χ) is defined on $\mathbb{R}^n \times (0,\infty)$ and is a weak solution in Ω_T for any $T > 0$, we call it a global weak solution of (1.11)–(1.14).

Note that by (2.17), $0 \le \theta(w,u) \le 1$. Hence (2.18) makes sense.

In the next sections, we will show that there exists a weak solution (w,u,χ) of (1.11)–(1.14). Here we point out that the u^2 in (2.13) has been replaced by χ in (2.18). This is technically necessary because we will encounter the weak convergence of some approximate sequence u_ε of u. It is well known that the weak limit χ of u_ε^2, if it ever exists, is not necessarily equal to u^2. From the above definition of weak solutions, it seems that the function χ is not well determined. However, the following result tells us that this χ cannot be arbitrary.

PROPOSITION 2.2. *Suppose that* $\lambda > 0$ *and* $\gamma > 0$. *Let* (w, u, χ) *and* $(\widehat{w}, \widehat{u}, \widehat{\chi})$ *be two weak solutions of* (1.11)–(1.14) *in the sense of Definition* 2.1. *Then the following hold:*

(i) *If* $w = \widehat{w}$, *then* $u = \widehat{u}$ *and* $\chi = \widehat{\chi}$.

(ii) *If* $u = \widehat{u}$, *then* $w = \widehat{w}$ *and* $\chi = \widehat{\chi}$.

Proof. (i) Let $w = \widehat{w}$. By (2.11) we obtain $u = \widehat{u}$. Then by (2.18) we have $\chi = \widehat{\chi}$. We can similarly prove (ii) by using (2.11) and (2.18). □

The above result implies that the function χ is uniquely determined by the pair (w, u) in the case in which $\lambda, \gamma > 0$. This, however, does not say that the weak solution is unique. In §4 we will give another alternative way of representing the function χ via the so-called Young measure. In §6 we will further show that for the one-dimensional case, if the initial data $w_0(x)$ has a positive lower bound, then

$$\chi(x,t) = u(x,t)^2. \tag{2.19}$$

Hence (w, u) actually satisfy (2.11) and (2.13) in this case.

3. Approximate problem. In this section we study the following approximate problem of (1.11)–(1.14): Let $T > 0$,

$$w_t = \nabla \cdot (w \nabla w) - \lambda w + \lambda E_\delta(w - \varphi_\varepsilon * u) \quad \text{in } \Omega_T, \tag{3.1}$$

$$u_t = \nabla u \cdot \nabla w + (\Delta w + \gamma w - \lambda - \beta)u - \gamma u^2 \quad \text{in } \Omega_T, \tag{3.2}$$

$$w\,|_{t=0} = w_0(x) + \varepsilon \equiv w_0^\varepsilon(x), \qquad x \in \mathbb{R}^n, \tag{3.3}$$

$$u\,|_{t=0} = u_0(x) + \varepsilon \equiv u_0^\varepsilon(x), \qquad x \in \mathbb{R}^n. \tag{3.4}$$

In the above, $0 < \varepsilon, \delta \le 1$; $E_\delta(s)$ is a nonnegative smooth function defined in $\mathbb{R}$ nondecreasingly converging to $E_0(s) \equiv \max\{s, 0\}$ as $\delta \downarrow 0$, and with $E_\delta'(s)$ being bounded independent of δ (this is possible since $E_0(s)$ is Lipschitz continuous); and φ_ε is an ε-mollifier

$$\varphi_\varepsilon(x,t) = \frac{1}{\varepsilon^{n+1}} \varphi\left(\frac{x}{\varepsilon}, \frac{t}{\varepsilon}\right),$$

with

$$\varphi \in C_0^\infty(\mathbb{R}^{n+1}), \quad \operatorname{supp}\varphi \subset \{(x,t) \in \mathbb{R}^{n+1} \mid |x|^2 + t^2 \le 1, t \ge 0\},$$
$$\varphi(x,t) \ge 0, \quad \int_{\mathbb{R}^{n+1}} \varphi(x,t)\,dx\,dt = 1.$$

We call problem (3.1)–(3.4) the approximate problem of (1.11)–(1.14).

We have seen that (3.2) is the same as (1.12). However, (1.11) is changed to (3.1). The purpose of making this approximation is that if w is a solution of (3.1), then $w \ge 0$. We do not a priori have such a property for a solution of (1.11).

The main result of this section is the following theorem.

THEOREM 3.1. *There exists a unique classical solution* $(w_{\varepsilon,\delta}, u_{\varepsilon,\delta})$ *of* (3.1)–(3.4) *for any* $T > 0$. *Moreover, this solution satisfies*

$$(\bar{w}_0 + \varepsilon)e^{-MT} \le w_{\varepsilon,\delta}(x,t) \le \|w_0\|_{C^0(\Omega)} + \varepsilon, \qquad (x,t) \in \bar{\Omega}_T, \tag{3.5}$$

$$0 < u_{\varepsilon,\delta}(x,t) \le w_{\varepsilon,\delta}(x,t) \quad \forall (x,t) \in \bar{\Omega}_T \quad \forall T > 0, \tag{3.6}$$

where $\bar{w}_0 = \inf_{x \in \mathbb{R}^n} w_0(x) \ge 0$.

To prove the above theorem we begin with the study of equation (3.1). In what follows we let $T_0 > 0$ be fixed and $0 < T \le T_0$. The following lemma was proved in [11].

LEMMA 3.2. *Let* $w \in C^0(\bar{\Omega}_T)$ *with* $D^2 w \in C^0(\bar{\Omega}_T)$ *and* $(\bar{x}, \bar{t}) \in \Omega_T$. *Then there exists a unique solution* $\psi(t; \bar{x}, \bar{t})$ *of*

$$\begin{cases} \dfrac{d\psi}{dt} = -\nabla w(\psi, t), & 0 < t < T, \\ \psi(\bar{t}; \bar{x}, \bar{t}) = \bar{x}. \end{cases} \tag{3.7}$$

Moreover, the solution $\psi(t; \bar{x}, \bar{t})$ *is differentiable with respect to all its arguments and satisfies*

$$\frac{\partial \psi(t; \bar{x}, \bar{t})}{\partial \bar{t}} = D_{\bar{x}} \psi(t; \bar{x}, \bar{t}) \nabla w(\bar{x}, \bar{t}), \qquad t \in [0, T], \tag{3.8}$$

$$\begin{cases} \dfrac{d}{dt} D_{\bar{x}} \psi(t; \bar{x}, \bar{t}) = -D^2 w(\psi(t; \bar{x}, \bar{t}), t) D_{\bar{x}} \psi(t; \bar{x}, \bar{t}), \\ D_{\bar{x}} \psi(\bar{t}; \bar{x}, \bar{t}) = I. \end{cases} \tag{3.9}$$

Consequently,

$$|D_{\bar{x}} \psi(t; \bar{x}, \bar{t})| \le e^{T \| D^2 w \|_{C^0(\bar{\Omega}_T)}}, \qquad t \in [0, T]. \tag{3.10}$$

We refer to the solution ψ of (3.7) as the characteristics associated with w. The following result gives a representation of solutions to (3.2) and (3.4).

LEMMA 3.3. *Let* $w \in C^0(\bar{\Omega}_T)$ *with* $D^2 w \in C^0(\bar{\Omega}_T)$. *Let* $\psi(t; \bar{x}, \bar{t})$ *be determined by* (3.7). *Let* u *be a classical solution of* (3.2) *and* (3.4) *in* Ω_T. *Then* u *can be expressed as*

$$u(\bar{x}, \bar{t}) = \frac{u_0^\varepsilon(\psi(0; \bar{x}, \bar{t})) e^{\int_0^{\bar{t}} h(\psi(t; \bar{x}, \bar{t}), t) dt}}{1 + u_0^\varepsilon(\psi(0; \bar{x}, \bar{t})) \int_0^{\bar{t}} \gamma(\psi(t; \bar{x}, \bar{t}), t) e^{\int_0^t h(\psi(\tau; \bar{x}, \bar{t}), \tau) d\tau} dt}, \tag{3.11}$$

where $h(x,t) = \Delta w(x,t) + \gamma(x,t) w(x,t) - \lambda(x,t) - \beta(x,t)$.

Proof. We fix $(\bar{x}, \bar{t}) \in \Omega_T$ and set

$$u(t) = u(\psi(t; \bar{x}, \bar{t}), t).$$

Then $u(t)$ satisfies the following:

$$\begin{cases} \dfrac{du(t)}{dt} = h(\psi(t; \bar{x}, \bar{t}), t) u(t) - \gamma u(t)^2, \\ \\ u(0) = u_0^\varepsilon(\psi(0; \bar{x}, \bar{t})) > 0. \end{cases} \tag{3.12}$$

It follows from (3.12) that $u(t) > 0$ for all $t \in [0,T]$. Thus $q(t) = \frac{1}{u(t)}$ is well defined and satisfies

$$\begin{cases} \dfrac{dq(t)}{dt} = -hq(t) + \gamma, \\ q(0) = \dfrac{1}{u(0)}. \end{cases} \tag{3.13}$$

Hence,

$$q(t) = e^{-\int_0^t h d\tau} q(0) + \int_0^t \gamma e^{-\int_s^t h d\tau} ds. \tag{3.14}$$

We thus obtain (3.11) by taking $t = \bar{t}$. □

Note that u can always be defined by (3.11) provided $\Delta w \in C^0(\bar{\Omega}_T)$. However, if w has no more regularity, this function may not necessarily be a classical solution of (3.2).

LEMMA 3.4. *Let $w \in C^0(\bar{\Omega}_T)$ with $\Delta w \in C^0(\bar{\Omega}_T)$. Let u be defined by* (3.11). *Then*

$$\|u\|_{C^0(\bar{\Omega}_T)} \le (1 + \|u_0\|_{C^0(\Omega)}) e^{T(\|\Delta w\|_{C^0(\bar{\Omega}_T)} + M\|w\|_{C^0(\bar{\Omega}_T)})}. \tag{3.15}$$

Proof. It is easy to see that (note (2.7))

$$e^{\int_0^{\bar{t}} (\Delta w + \gamma w - \lambda - \beta) dt} \le e^{T(\|\Delta w\|_{C^0(\bar{\Omega}_T)} + M\|w\|_{C^0(\bar{\Omega}_T)})}, \qquad \bar{t} \in [0,T]. \tag{3.16}$$

Hence (3.15) follows from (3.11) (since $0 < \varepsilon \le 1$). □

LEMMA 3.5. *If $u \in C^0(\bar{\Omega}_T)$ and $u(x,t) \ge 0$ on $\bar{\Omega}_T$, then* (3.1) *and* (3.3) *admit a unique solution $w \in C^{2,\alpha}(\bar{\Omega}_T)$ with $\alpha \in (0,1)$ only depending on $\varepsilon > 0$. Moreover, this solution satisfies* (3.5).

Proof. For any $\sigma > 0$, we let $G_\sigma(s) \in C^\infty$ be a function satisfying

$$G_\sigma(s) = \begin{cases} s, & s \ge \sigma, \\ \sigma/2, & s \le \sigma/2 \end{cases} \tag{3.17}$$

and $G_\sigma(s) \ge \sigma/2$. Since E_δ is C^1 with E'_δ bounded uniformly in $\delta > 0$, by the standard results for the nonlinear parabolic equations [12], for any $u \in C^0(\bar{\Omega}_T)$, the Cauchy problem

$$\begin{cases} \bar{w}_t - \nabla \cdot [G_\sigma(\bar{w}) \nabla \bar{w}] = -\lambda \bar{w} + \lambda E_\delta(\bar{w} - \varphi_\varepsilon * u), \\ \bar{w}(0,x) = w_0^\varepsilon(x) \end{cases} \tag{3.18}$$

is uniquely solvable in $C^{2,\alpha}(\bar{\Omega}_T)$. On the other hand, since $E_\delta \ge 0$, we get

$$\bar{w}_t - \nabla \cdot [G_\sigma(\bar{w}) \nabla \bar{w}] + \lambda \bar{w} \ge 0. \tag{3.19}$$

By the maximum principle we thus obtain

$$\bar{w}(x,t) \ge (\bar{w}_0 + \varepsilon) e^{-MT} \quad \forall (x,t) \in \bar{\Omega}_T. \tag{3.20}$$

Therefore, if we choose $\sigma \leq \varepsilon e^{-MT}$ at the beginning, then $\bar{w}$ is actually the unique solution of (3.1) and (3.3). Now we rewrite equation (3.1) as the following:

$$\bar{w}_t - \nabla \cdot (\bar{w} \nabla \bar{w}) + c(x,t)\bar{w} = 0, \tag{3.21}$$

where

$$c(x,t) = \lambda - \frac{\lambda E_\delta(\bar{w} - \varphi_\varepsilon * u)}{\bar{w}}. \tag{3.22}$$

By (3.20) we know that $\bar{w} > 0$. Thus, note that $\varphi_\varepsilon * u \geq 0$ and $E_\delta \uparrow E_0$; we have

$$0 \leq \frac{E_\delta(\bar{w} - \varphi_\varepsilon * u)}{\bar{w}} \leq \frac{E_0(\bar{w} - \varphi_\varepsilon * u)}{\bar{w}} \leq 1. \tag{3.23}$$

This yields that

$$0 \leq c(x,t) \leq M, \qquad (x,t) \in \bar{\Omega}_T. \tag{3.24}$$

Hence, by maximum principle we obtain

$$\bar{w}(x,t) \leq \|w_0\|_{C^0(\Omega)} + \varepsilon. \tag{3.25}$$

Combining (3.20) and (3.25), we obtain (3.5). □

We now construct a mapping $\mathcal{A} : C^{2,0}(\bar{\Omega}_T) \to C^{2,\alpha}(\bar{\Omega}_T) \subset C^{2,0}(\bar{\Omega}_T)$ as follows: For any $w \in C^{2,0}(\bar{\Omega}_T)$, let $u \in C^0(\bar{\Omega}_T)$ be defined by (3.11). We designate $\bar{w} = \mathcal{A}w$ as the unique solution of (3.1) and (3.3). From the above arguments $\mathcal{A}$ is well defined and satisfies (3.5). We have the following lemma about this map.

LEMMA 3.6. *There exists a $T_1 \in (0,T]$ such that the map $\mathcal{A} : C^{2,0}(\bar{\Omega}_{T_1}) \to C^{2,0}(\bar{\Omega}_{T_1})$ admits a fixed point $w \equiv w_{\varepsilon,\delta} \in C^{2,\alpha}(\bar{\Omega}_T)$.*

Proof. Consider the convex and closed set

$$B = \{w \in C^{2,0}(\bar{\Omega}_{T_1}) \mid \|w\|_{C^{2,0}(\bar{\Omega}_{T_1})} \leq K\},$$

where $0 < T_1 \leq T$ and $K > 0$ are undetermined. From (3.21) and (3.24), by maximum principle, the L^p-estimates, and the Sobolev inequalities we get

$$\|\bar{w}\|_{C^{1,\alpha}(\bar{\Omega}_{T_1})} \leq C(\varepsilon,T). \tag{3.26}$$

Here, the constant $C(\varepsilon,T)$ depends on ε because of (3.20). Then, applying the Schauder estimates, we derive (note (3.15))

$$\begin{aligned} \|\bar{w}\|_{C^{2,\alpha}(\bar{\Omega}_{T_1})} &\leq C(\varepsilon,T)[1 + \|w_0\|_{C^{2+\alpha}(\Omega)} + \|\varphi_\varepsilon * u\|_{C^\alpha(\bar{\Omega}_{T_1})}] \\ &\leq C(\varepsilon,T)\left[1 + \|w_0\|_{C^{2+\alpha}(\Omega)} + \frac{C}{\varepsilon}\|u\|_{C^0(\bar{\Omega}_{T_1})}\right] \\ &\leq C(\varepsilon,T)\left[1 + \|w_0\|_{C^{2+\alpha}(\Omega)} + \frac{C}{\varepsilon}(1 + \|u_0\|_{C^0(\Omega)})e^{T_1K(1+M)}\right]. \end{aligned} \tag{3.27}$$

We now take

$$\begin{aligned} K &= C(\varepsilon,T)\left[1 + \|w_0\|_{C^{2+\alpha}(\Omega)} + \frac{C}{\varepsilon}e(1 + \|u_0\|_{C^0(\Omega)})\right], \\ T_1 &= \frac{1}{K(1+M)} \equiv T_1(\varepsilon,T,\|w_0\|_{C^{2+\alpha}(\Omega)},\|u_0\|_{C^0(\Omega)}). \end{aligned} \tag{3.28}$$

Then, from (3.27) we see that $\mathcal{A}$ maps B into itself. It is clear that this map is continuous and compact. By the Schauder fixed point theorem $\mathcal{A}$ possesses a fixed point. □

LEMMA 3.7. *Let $w \equiv w_{\varepsilon,\delta}$ be a fixed point of the map $\mathcal{A}$ in $C^{2,\alpha}(\bar{\Omega}_{T_1})$ and let $u \equiv u_{\varepsilon,\delta}$ be defined by* (3.11) *with $\psi \equiv \psi_{\varepsilon,\delta}$ being the characteristic associated with $w_{\varepsilon,\delta}$. Then $u \in C^1(\bar{\Omega}_{T_1})$ and $(w,u) \equiv (w_{\varepsilon,\delta}, u_{\varepsilon,\delta})$ is a classical solution of* (3.1)–(3.4) *on Ω_{T_1} satisfying* (3.5)–(3.6).

Proof. To show that (w,u) is a classical solution of (3.1)–(3.4) on Ω_{T_1}, we only need to verify (3.2). Since w satisfies (3.1), we get

$$\Delta w(\psi(t;\bar{x},\bar{t}),t) = \frac{1}{w(\psi(t;\bar{x},\bar{t}),t)} \frac{dw(\psi(t;\bar{x},\bar{t}),t)}{dt} + c(\psi(t;\bar{x},\bar{t}),t), \tag{3.29}$$

where $\psi(\cdot;\bar{x},\bar{t})$ is the characteristic associated with $w \equiv w_{\varepsilon,\delta}$ (see (3.7)) and $c(\cdot)$ is defined by (3.22) with $(\bar{w},u) = (w,u) \equiv (w_{\varepsilon,\delta}, u_{\varepsilon,\delta})$. Hence (still let $h(x,t) = \Delta w + \gamma w - \lambda - \beta$)

$$\begin{aligned} g(t;\bar{x},\bar{t}) &\equiv e^{\int_0^t h(\psi(\tau;\bar{x},\bar{t}),\tau)d\tau} = e^{\int_0^t \frac{1}{w}\frac{dw}{dt}d\tau} \cdot e^{\int_0^t (c+\gamma w-\lambda-\beta)d\tau} \\ &= \frac{w(\psi(t;\bar{x},\bar{t}),t)}{w_0^\varepsilon(\psi(0;\bar{x},\bar{t}))} e^{\int_0^t (c+\gamma w-\lambda-\beta)d\tau}. \end{aligned} \tag{3.30}$$

Clearly, the right-hand side of (3.30) belongs to C^1 (as a function of $(t,\bar{x},\bar{t})$). By (3.11), the definition of u, and (3.30), we have

$$u(\bar{x},\bar{t}) = \frac{u_0^\varepsilon(\psi(0;\bar{x},\bar{t}))g(\bar{t};\bar{x},\bar{t})}{1 + u_0^\varepsilon(\psi(0;\bar{x},\bar{t})) \int_0^{\bar{t}} \gamma(\psi(t;\bar{x},\bar{t})g(t;\bar{x},\bar{t})dt}. \tag{3.31}$$

Therefore $u \in C^1$ (see (3.8)–(3.10)). Furthermore, by (3.8) we derive

$$\begin{aligned} \frac{\partial u_0^\varepsilon}{\partial \bar{t}}(\psi(0;\bar{x},\bar{t})) &= \nabla_x u_0^\varepsilon(\psi(0;\bar{x},\bar{t}))D_{\bar{x}}\psi(0;\bar{x},\bar{t})\nabla_{\bar{x}} w(\bar{x},\bar{t}) \\ &= \nabla_{\bar{x}} u_0^\varepsilon(\psi(0;\bar{x},\bar{t})) \cdot \nabla_{\bar{x}} w(\bar{x},\bar{t}). \end{aligned} \tag{3.32}$$

Analogously, we have

$$\frac{\partial g}{\partial \bar{t}}(t;\bar{x},\bar{t}) = \nabla_{\bar{x}} g(t;\bar{x},\bar{t}) \cdot \nabla_{\bar{x}} w(\bar{x},\bar{t}) \tag{3.33}$$

and

$$\begin{aligned} \frac{\partial g}{\partial t}(t;\bar{x},\bar{t}) &= h(\psi(t;\bar{x},\bar{t}),t)g(t;\bar{x},\bar{t}) \\ &= (\Delta w + \gamma w - \lambda - \beta)g. \end{aligned} \tag{3.34}$$

Then, by direct computation, we are able to show that u satisfies (3.2).

Next, from (3.11), $u > 0$ on Ω_{T_1}. Then, by Lemma 3.5 we know that (3.5) holds. Finally, we set $v = w - u$. Then v solves

$$\begin{cases} v_t - \nabla v \cdot \nabla w - (\Delta w - \lambda - \gamma u)v = \beta u + \lambda E_\delta(w - \varphi_\varepsilon * u) \geq 0, \\ v\,|_{t=0} = w_0 - u_0 = v_0 \geq 0. \end{cases} \tag{3.35}$$

By integrating v along the characteristics (3.7), we obtain $v \geq 0$. Hence $u \leq w$. This completes the proof. □

Now, we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. By Lemma 3.7 we know that there exists a $T_1 \leq T$ such that (3.1)–(3.4) has a classical solution on $\bar{\Omega}_{T_1}$. Let us show that the classical solution on Ω_{T_1} is unique. To this end we suppose there are two solutions (w_1, u_1) and (w_2, u_2) in Ω_{T_1}. Let $\widetilde{u} = u_1 - u_2$ and $\widetilde{w} = w_2 - w_2$. Then

$$\widetilde{w}_t - \frac{1}{2}(w_1 + w_2)\Delta\widetilde{w} = \nabla(w_1 + w_2) \cdot \nabla\widetilde{w} + a_1\widetilde{w} - a_2\varphi_\varepsilon * \widetilde{u}, \tag{3.36}$$

$$\widetilde{u}_t - \nabla\widetilde{u} \cdot \nabla w_1 = b_1\widetilde{u} + b_2, \tag{3.37}$$

$$\widetilde{w}\,|_{t=0} = 0, \tag{3.38}$$

$$\widetilde{u}\,|_{t=0} = 0, \tag{3.39}$$

where

$$\begin{aligned} a_1 &= \frac{1}{2}\Delta(w_1 + w_2) - \lambda + a_2, \\ a_2 &= \lambda \int_0^1 E_\delta'(\tau(\widetilde{w} - \varphi_\varepsilon * \widetilde{u}) + w_2 - \varphi_\varepsilon * u_2)d\tau, \\ b_1 &= \Delta w_1 + \gamma w_1 - \lambda - \beta - \gamma(u_1 + u_2), \\ b_2 &= \nabla u_2 \cdot \nabla\widetilde{w} + (\Delta\widetilde{w} + \gamma\widetilde{w})u_2. \end{aligned} \tag{3.40}$$

Since $w_1, w_2 \in C^{2,\alpha}(\bar{\Omega}_{T_1})$ and $u_1, u_2 \in C^1(\bar{\Omega}_{T_1})$, we have

$$\begin{aligned} &\|a_1\|_{C^{0,\alpha}(\bar{\Omega}_{T_1})}, \|a_2\|_{C^{0,\alpha}(\bar{\Omega}_{T_1})}, \|b_1\|_{C^{0,\alpha}(\bar{\Omega}_{T_1})} \leq C, \\ &\|b_2\|_{C^0} \leq C\|D^2\widetilde{w}\|_{C^0(\bar{\Omega}_{T_1})}. \end{aligned} \tag{3.41}$$

Integrating (3.37) along the characteristics corresponding to w_1, we get

$$\widetilde{u}(\bar{x}, \bar{t}) = \int_0^{\bar{t}} e^{\int_t^{\bar{t}} b_1(\psi(\tau;\bar{x},\bar{t}),\tau)d\tau} b_2 dt. \tag{3.42}$$

Hence, for any $\widetilde{T}_1 \leq T_1$ we have

$$|\widetilde{u}(\bar{x}, \bar{t})| \leq \bar{t}C\|b_2\|_{C^0(\bar{\Omega}_{\widetilde{T}_1})} \leq \bar{t}C\|D^2\widetilde{w}\|_{C^0(\bar{\Omega}_{\widetilde{T}_1})} \quad \forall \bar{t} \leq \widetilde{T}_1. \tag{3.43}$$

On the other hand, by Schauder estimates we get from (3.36) that

$$\|\widetilde{w}\|_{C^{2+\alpha}(\bar{\Omega}_{\widetilde{T}_1})} \leq C\|\varphi_\varepsilon * \widetilde{u}\|_{C^\alpha(\bar{\Omega}_{\widetilde{T}_1})} \leq C_\varepsilon\|\widetilde{u}\|_{C^0(\bar{\Omega}_{\widetilde{T}_1})}. \tag{3.44}$$

Substituting (3.44) into (3.43), we obtain that in $\Omega_{\widetilde{T}_1}$ $(\widetilde{T}_1 \leq T_1)$,

$$\|\widetilde{u}\|_{C^0(\bar{\Omega}_{\widetilde{T}_1})} \leq \widetilde{T}_1\widetilde{C}_\varepsilon\|\widetilde{u}\|_{C^0(\bar{\Omega}_{\widetilde{T}_1})}. \tag{3.45}$$

Note that $\widetilde{C}_\varepsilon$ does not depend on $\widetilde{T}_1$. Hence it follows that $\widetilde{u} = 0$ in $\bar{\Omega}_{\widetilde{T}_1}$ if $\widetilde{T}_1 \le (2\widetilde{C}_\varepsilon)^{-1}$. Then $\widetilde{w} = 0$ in the same region. Repeating this procedure, we can get $\widetilde{u} = \widetilde{w} = 0$ in Ω_{T_1}. This shows the uniqueness of the local classical solutions of (3.1)–(3.4).

Next we prove the existence of global classical solutions. To this end we only need to derive some a priori estimate. Suppose (w, u) is a classical solution of (3.1)–(3.4) satisfying (3.5)–(3.6). Then, for any $\bar{T} \le T$, similar to (3.27), we have

$$
\begin{aligned}
\|w\|_{C^{2,\alpha}(\bar{\Omega}_{\bar{T}})} &\le C(\varepsilon, T)[1 + \|w_0\|_{C^{2+\alpha}(\Omega)} + \|\varphi_\varepsilon * u\|_{C^\alpha(\bar{\Omega}_{\bar{T}})}] \\
&\le C(\varepsilon, T)\left[1 + \|w_0\|_{C^{2+\alpha}(\Omega)} + \frac{C}{\varepsilon}\|u\|_{C^0(\bar{\Omega}_{\bar{T}})}\right] \\
&\le C(\varepsilon, T)\left[1 + \|w_0\|_{C^{2+\alpha}(\Omega)} + \frac{C}{\varepsilon}(\|w_0\|_{C^0(\Omega)} + \varepsilon)\right] \\
&\le \bar{C}(\varepsilon, T)(1 + \|w_0\|_{C^{2+\alpha}(\Omega)}).
\end{aligned} \tag{3.46}
$$

In (3.46), we have used (3.5)–(3.6). Since the above estimate is uniform in $\bar{T} \in [0, T]$, we can repeat the procedure given in the proof of Lemma 3.6 to extend the solution (w, u) to Ω_T. Since for any $T > 0$ there exists a unique classical solution (w, u) in Ω_T, the solution can be extended to $\Omega \times (0, \infty)$ to get a global solution. □

4. Existence of weak solutions. In this section we show the existence of a weak solution to (1.11)–(1.14). From the previous section we know that for any $\delta, \varepsilon > 0$, there exists a unique classical solution $(w_{\delta,\varepsilon}, u_{\delta,\varepsilon})$ of the approximate problem (3.1)–(3.4). We first study the limit as $\delta \to 0$.

LEMMA 4.1. *For any $\varepsilon > 0$ there exists a classical solution $(w_\varepsilon, u_\varepsilon)$ of*

$$
w_t - \nabla \cdot [w \nabla w] = -\lambda w + \lambda E_0(w - \varphi_\varepsilon * u) \tag{4.1}
$$

and (3.2)–(3.4) *such that*

$$
0 < u_\varepsilon \le w_\varepsilon \le \|w_0\|_{C^0(\Omega)}, \qquad w_\varepsilon(x, t) \ge (\bar{w}_0 + \varepsilon)e^{-Mt}. \tag{4.2}
$$

Proof. For fixed $T > 0$, from (3.5)–(3.6) we know that $w_{\delta,\varepsilon}$ and $u_{\delta,\varepsilon}$ are bounded uniformly with respect to $\delta > 0$. By the Schauder estimates (see [12]) we have

$$
\|w_{\varepsilon,\delta}\|_{C^{2,\alpha}(\bar{\Omega}_T)} \le C(\varepsilon, T). \tag{4.3}
$$

It follows from (3.11) that

$$
\|u_{\varepsilon,\delta}\|_{C^1(\bar{\Omega}_T)} \le C(\varepsilon, T). \tag{4.4}
$$

Hence we can select a subsequence, still denoted by itself, and $w_\varepsilon \in C^{2,\alpha}(\bar{\Omega}_T)$, u_ε being Lipschitz, such that for some $\mu \in (0, \alpha)$,

$$
w_{\varepsilon,\delta} \to w_\varepsilon \quad \text{in } C^{2,\mu}(\bar{\Omega}_T), \tag{4.5}
$$

$$
u_{\varepsilon,\delta} \to u_\varepsilon \quad \text{in } C^{\mu}(\bar{\Omega}_T) \tag{4.6}
$$

as $\delta \to 0$. By the diagonal argument we can further select a subsequence and $w_\varepsilon \in C^{2,\alpha}(\mathbb{R}^+ \times \mathbb{R}^n)$, u being Lipschitz in $\mathbb{R}^n \times \mathbb{R}^+$, such that (4.1) and (4.2) hold in

Ω_T for any $T > 0$. It is clear that $(w_\varepsilon, u_\varepsilon)$ satisfies (4.1). Since the characteristic (3.7) depends continuously on $D^2 w_{\delta,\varepsilon}$, we can pass to the limits through (3.11). It follows that the limit u_ε has the same expression (3.11). From the proof of Lemma 3.7, $u_\varepsilon \in C^1$ and $(w_\varepsilon, u_\varepsilon)$ satisfies (3.2). Finally, (4.2) follows from (3.5)–(3.6) and (3.11). □

Next, we look at the variational forms of the ε-approximate solution $(w_\varepsilon, u_\varepsilon)$. It is easy to see that for any $\xi, \eta \in C_0^\infty(\bar{\Omega}_T^0)$, $(w_\varepsilon, u_\varepsilon)$ satisfy the following:

$$\begin{aligned} &\int_{\Omega_T} \left\{ w_\varepsilon \xi_t - \frac{1}{2} \nabla w_\varepsilon^2 \cdot \nabla \xi - \lambda w_\varepsilon \xi + \lambda E_0 (w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) \xi \right\} dxdt \\ &\qquad = \int_{\mathbf{R}^n} (w_0(x) + \varepsilon) \xi(x, 0) dx, \end{aligned} \tag{4.7}$$

$$\begin{aligned} &\int_{\Omega_T} \left\{ u_\varepsilon \eta_t - \frac{1}{2} \theta(w_\varepsilon, u_\varepsilon) \nabla w_\varepsilon^2 \cdot \nabla \eta + [(\gamma w_\varepsilon - \lambda - \beta) u_\varepsilon - \gamma u_\varepsilon^2] \eta \right\} dxdt \\ &\qquad = \int_{\mathbf{R}^n} (u_0(x) + \varepsilon) \eta(x, 0) dx, \end{aligned} \tag{4.8}$$

where θ is defined by (2.14). We call $(w_\varepsilon, u_\varepsilon)$ the ε-approximate solution.

LEMMA 4.2. *Let $(w_\varepsilon, u_\varepsilon)$ be as in Lemma 4.1 and $\Omega_T(R) = B_R(0) \times (0, T)$, where $B_R(0)$ is the ball in $\mathbb{R}^n$ centered at the origin with radius R. Then there exists a constant $C(R)$, depending only on R, such that*

$$\int_{\Omega_T(R)} |\nabla w_\varepsilon^2|^2 dxdt \le C(R) \quad \forall \varepsilon > 0. \tag{4.9}$$

Proof. In (4.7) and (4.8) we put $2T$ in place of T and $w_\varepsilon^2 \xi^2$ in place of ξ with $\xi \in C^\infty(\mathbb{R}^n \times \mathbb{R}^+)$, $0 \le \xi \le 1$, $\xi = 1$ on $\Omega_T(R)$, and $\xi = 0$ outside of $\Omega_{2T}(2R)$. It follows that

$$\begin{aligned} &\int_{\Omega_{2T}} |\nabla w_\varepsilon^2|^2 \xi^2 dxdt \\ &\le \int_{\Omega_{2T}} \left\{ w_\varepsilon^2 |\nabla w_\varepsilon^2| |\nabla \xi|^2 + 2[\lambda w_\varepsilon^3 + \lambda E_0 (w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) w_\varepsilon^2] \xi^2 \right\} dxdt \\ &+ \left| \int_{\Omega_{2T}} 4 w_\varepsilon^2 (w_\varepsilon)_t \xi^2 dxdt \right| + \left| \int_{\Omega_{2T}} 4 w_\varepsilon^2 \xi \xi_t dxdt \right| + \int_\Omega (w_0 + \varepsilon)^3 \xi(x, 0)^2 dx. \end{aligned} \tag{4.10}$$

We have

$$\begin{aligned} \int_{\Omega_{2T}} w_\varepsilon^2 (w_\varepsilon)_t \xi^2 dxdt &= \int_{\Omega_{2T}} \frac{1}{3} (w_\varepsilon^3)_t \xi^2 dxdt \\ &= -\frac{2}{3} \int_{\Omega_{2T}} \xi \xi_t w_\varepsilon^3 dxdt - \frac{1}{3} \int_{\mathbf{R}^n} (w_0 + \varepsilon)^3 \xi(x, 0) dx. \end{aligned}$$

Since w_ε and u_ε are uniformly bounded, (4.9) follows from (4.10). □

LEMMA 4.3. *Let $(w_\varepsilon, u_\varepsilon)$ be the ε-approximate solution. Then there exist $w \in C^{0,\alpha}(\bar{\Omega}_T)$, $u \in L^\infty(\Omega_T)$ such that for any bounded open set $Q \subset \mathbb{R}^n \times \mathbb{R}^+$,*

$$u_\varepsilon \to u, \quad \text{weakly in } L^2(Q), \tag{4.11}$$

$$\nabla w_\varepsilon^2 \to \nabla w^2, \qquad \text{weakly in } L^2(Q), \tag{4.12}$$

$$w_\varepsilon \to w \quad \text{strongly in } C^{0,\alpha}(Q). \tag{4.13}$$

Proof. For fixed $R > 0$ and $T > 0$, from Lemma 4.2 and the previous section w_ε, ∇w_ε^2, and u_ε are bounded in $L^2(\Omega_T(R))$. Hence we can select a sequence such that (4.11), (4.12) hold in $\Omega_T(R)$. By [5] and [6], we know that there exists a subsequence such that (4.13) holds in $\Omega_T(R)$. Then the assertion of the lemma follows from the diagonalization argument. □

We now can pass to the limits in (4.7). The limit w is a weak solution of the porous medium equation

$$w_t - \nabla \cdot [w \nabla w] = -\lambda w + \lambda G, \tag{4.14}$$

where G is the weak limit of $E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon)$. Set

$$\Omega_T^+ = \{(x,t) \in \Omega_T : w(x,t) > 0\};$$

Γ_T is the boundary of Ω_T^+. Since w is continuous, Ω_T^+ is open. Let $\Omega^t = \{x \in \mathbb{R}^n : w(x,t) > 0\}$. To study the properties of Ω^t, we need [10, Lem. 3.6]. For the reader's convenience, we state that result here.

LEMMA 4.4. *Let $g(x,t)$ be a bounded function in Ω_T; w is a weak solution of*

$$\begin{cases} w_t - \nabla \cdot (w \nabla w) = g(x,t) w & \text{in } \Omega_T, \\ w(x,0) = w_0(x) \geq 0. \end{cases} \tag{4.15}$$

Let Ω^t and Γ_T be defined as above. Then Ω^t is increasing in t and the $(n+1)$-dimensional Lebesgue measure of Γ_T is zero.

For our problem we have the following result.

LEMMA 4.5. *Suppose $w_0(x)$ has a compact support and (2.10) holds. Let (w,u) be a limit of ε-approximate solutions $(w_\varepsilon, u_\varepsilon)$ in the sense of (4.11)–(4.13). Then the region Ω^t is increasing in t and the boundary Γ_T of Ω_T^+ has the $(n+1)$-dimensional Lebesgue measure zero.*

Proof. Since $0 \leq E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) \leq w_\varepsilon$, we have $0 \leq G \leq w$. Let

$$g = -\lambda + \lambda \frac{G}{w}. \tag{4.16}$$

Then g is bounded and w satisfies

$$w_t - \nabla \cdot (w \nabla w) = gw. \tag{4.17}$$

Thus the assertion follows from Lemma 4.4. □

THEOREM 4.6. *There exists a weak solution (w,u,χ) of the problem (1.11)–(1.14). Moreover, $w \in C^{0,\alpha}(\bar{\Omega}_T)$ and $\nabla w \in C^{0,\alpha}_{\text{loc}}(\Omega_T^+)$.*

Proof. From Lemma 4.3 we can select a subsequence such that (4.11)–(4.13) hold. For any $\xi \in C_0^\infty(\bar{\Omega}_T^0)$ we write

$$\begin{aligned} &\int_{\Omega_T} \lambda E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon)\xi dx dt \\ &= \int_{\Omega_T} \lambda (w_\varepsilon - \varphi_\varepsilon * u_\varepsilon)\xi dx dt + \int_{\Omega_T} \lambda [E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) - (w_\varepsilon - \varphi_\varepsilon * u_\varepsilon)]\xi dx dt. \end{aligned} \tag{4.18}$$

Since $u_\varepsilon \to u$ weakly in $L^2_{\rm loc}(\Omega_T)$, it is clear that $\varphi_\varepsilon * u_\varepsilon \to u$ weakly in $L^2_{\rm loc}(\Omega_T)$. Hence

$$\int_{\Omega_T} \lambda(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon)\xi dxdt \to \int_{\Omega_T} \lambda(w-u)\xi dxdt. \tag{4.19}$$

By the definition

$$E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) - (w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) = \begin{cases} 0 & \text{if } w_\varepsilon - \varphi_\varepsilon * u_\varepsilon \geq 0, \\ -(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) & \text{if } w_\varepsilon - \varphi_\varepsilon * u_\varepsilon < 0. \end{cases} \tag{4.20}$$

On the other hand, by (4.2),

$$\varphi_\varepsilon * u_\varepsilon - w_\varepsilon \leq \varphi_\varepsilon * w_\varepsilon - w_\varepsilon. \tag{4.21}$$

It follows that for any $(x,t) \in \Omega_T$,

$$0 \leq E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) - (w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) \leq |\varphi_\varepsilon * w_\varepsilon - w_\varepsilon| \to 0, \tag{4.22}$$

since $w_\varepsilon \to w$ in $C^{0,\alpha}(\bar\Omega_T)$ locally. Hence

$$\int_{\Omega_T} \lambda[E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) - (w_\varepsilon - \varphi_\varepsilon * u_\varepsilon)]\xi dxdt \to 0. \tag{4.23}$$

This yields that

$$E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon) \to \lambda(w-u) \quad \text{weakly in } L^2_{\rm loc}(\Omega_T). \tag{4.24}$$

Thus (2.11) follows by sending $\varepsilon \to 0$ in (4.7).

In Ω_T^+, $w > 0$. Applying the $C^{1,\alpha}$-estimates [7], [12] in (4.1), we find that in any bounded open set Q with $\bar Q \subset \Omega_T^+$, $\|w_\varepsilon\|_{C^{1,\alpha}(Q)}$ is uniformly bounded. Hence, after extracting a subsequence, $\nabla w_\varepsilon \to \nabla w$ everywhere in Ω_T^+. In the case in which (2.9) holds, $\Omega_T^+ = \Omega_T$. Therefore the proof is complete. We now suppose that $w_0(x)$ has a compact support and (2.10) holds (case (2.9) is simpler). Then, for any point $(x_0,t_0) \notin \bar\Omega_T^+$, let $B_{2r}(x_0,t_0) \subset \Omega_T \setminus \bar\Omega_T^+$ be a ball centered at (x_0,t_0) with $r>0$ small enough. We choose a smooth ξ in (4.10) such that $\xi = 1$ in $B_r(x_0,t_0)$, $0 \leq \xi \leq 1$ and $\xi = 0$ outside of $B_{2r}(x_0,t_0)$. From (4.10), we get

$$\int_{B_r(x_0,t_0)} |\nabla w_\varepsilon^2|^2 dxdt \leq C \sup_{(x,t)\in B_{2r}(x_0,t_0)} |w_\varepsilon(x,t)| \to 0, \qquad \varepsilon \to 0. \tag{4.25}$$

Therefore

$$\nabla w_\varepsilon^2 \to 0 \quad \text{a.e. } (x,t) \in \Omega_T \setminus \bar\Omega_T^+. \tag{4.26}$$

By Lemma 4.5 we know that Γ_T has measure zero. Thus $\nabla w_\varepsilon^2 \to \nabla w^2$ almost everywhere in Ω_T. Finally, since $0 \leq u_\varepsilon \leq w_\varepsilon$ and $w_\varepsilon \to w$, we have

$$\theta(w_\varepsilon, u_\varepsilon) \to \theta(w,u) \quad \text{weakly in } L^2(\Omega_T). \tag{4.27}$$

Let χ be the weak limit of u_ε^2. We now pass to the limit in (4.8) to obtain (2.18). The proof is completed. □

To conclude this section let us give another way of expressing χ using the so-called Young measure. The result is stated as follows.

PROPOSITION 4.7. *There exists a parameterized probability measure $\nu(x,t,dr)$ supported on $[0,w(x,t)]$ such that*

$$u(x,t) = \int_0^{w(x,t)} r\nu(x,t,dr) \quad a.e.\ (x,t) \in \Omega \times (0,\infty), \tag{4.28}$$

$$\chi(x,t) = \int_0^{w(x,t)} r^2\nu(x,t,dr) \quad a.e.\ (x,t) \in \Omega \times (0,\infty). \tag{4.29}$$

Proof. We choose the sequence u_ε such that (4.11) holds. Then, as in [16], we can find a parameterized probability measure $\nu(x,t,dr)$ supported on $[0,w(x,t)]$ such that

$$F(u_\varepsilon^2(x,t)) \overset{w}{\rightharpoonup} \int_0^{w(x,t)} F(r)\nu(x,t,dr) \quad \text{in } L^2(Q) \tag{4.30}$$

for any continuous function F and any bounded open set $Q \subset \Omega \times (0,\infty)$. Thus, we obtain (4.28) and (4.29). □

We see that the above result gives a relation between u and χ. As soon as the parameterized measure $\nu(x,t,dr)$ is determined, the functions u and χ are automatically determined.

5. Existence of weak infective age distributions. In the previous section we proved the existence of weak solutions (w,u,χ) of (1.11)–(1.14). In particular we found a density u of the infectives. From our original model we also need to find an age distribution $\rho(x,t,a)$ of the infectives. The purpose of this section is to determine such a distribution. Hereafter we let $\mathcal{M}(0,\infty)$ be the set of all finite Borel measures on $[0,\infty)$. The spaces $L^p(\Omega_T;\mathcal{M}(0,\infty))$, $(1 \le p \le \infty)$ are defined in an obvious way.

For convenience let us rewrite equation (1.3) and conditions (1.5), (1.6) for ρ in the present case (i.e., λ, β, and γ are independent of a, and $v = w - u$):

$$\rho_t + \rho_a = \nabla \cdot [\rho \nabla w] - \lambda(x,t)\rho - \beta(x,t)\rho, \qquad (x,t,a) \in \mathbb{R}^n \times \mathbb{R}^+ \times \mathbb{R}^+, \tag{5.1}$$

$$\rho\,|_{t=0} = \rho_0(x,a), \qquad (x,a) \in \mathbb{R}^n \times \mathbb{R}^+, \tag{5.2}$$

$$\rho\,|_{a=0} = \gamma(x,t)[u(x,t)w(x,t) - u(x,t)^2], \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}^+. \tag{5.3}$$

We observe that if all the data involved in (5.1)–(5.3) are smooth and ρ is a classical solution of (5.1)–(5.3) smooth up to the boundary $\{t=0\} \cup \{a=0\}$, then the following compatibility condition should satisfy:

$$\begin{aligned} \rho_0(x,0) &= \gamma(x,0)[u_0(x)w_0(x) - u_0(x)^2] \\ &= \gamma(x,0)v_0(x)\int_0^\infty \rho_0(x,a)da, \qquad x \in \mathbb{R}^n. \end{aligned} \tag{5.4}$$

Here, we should note that $\rho_0(x,a)$ and $v_0(x)$ are given functions. Also, it is reasonable to assume that

$$\begin{aligned}&\rho_0 \in L^\infty(\mathbb{R}^n \times \mathbb{R}^+) \bigcap C^1(\mathbb{R}^n \times \mathbb{R}^+),\\ &\rho_0(x,a) \geq 0, \qquad (x,a) \in \mathbb{R}^n \times \mathbb{R}^+,\\ &\int_0^\infty \rho_0(x,a)da \equiv u_0(x) \in C^0(\mathbb{R}^n) \bigcap L^\infty(\mathbb{R}^n).\end{aligned} \tag{5.5}$$

In what follows, we will keep assumptions (5.4) and (5.5). On the other hand, if an age distribution $\rho(x,t,a)$ exists on Ω_T satisfying (5.1)–(5.3) and (1.1), then for any $\zeta \in C_0^\infty(\bar{\Omega}_T^0 \times [0,\infty))$ we have

$$\begin{aligned}&\int_{\Omega_T \times [0,\infty)} \rho\{-\zeta_t - \zeta_a + \nabla w \cdot \nabla \zeta + (\lambda+\beta)\zeta\}dxdtda\\ &+ \int_{\Omega \times [0,\infty)} \rho_0 \zeta(x,0,a)dxda + \int_{\Omega_T} \gamma(uw - u^2)\zeta(x,t,0)dxdt = 0.\end{aligned} \tag{5.6}$$

As in §2 we may also write (5.6) as follows:

$$\begin{aligned}&\int_{\Omega_T \times [0,\infty)} \Big\{\rho[-\zeta_t - \zeta_a + (\lambda+\beta)\zeta] + \frac{1}{2}\nabla w^2 \cdot \nabla \zeta \theta(w,\rho)\Big\}dxdtda\\ &+ \int_{\Omega \times [0,\infty)} \rho_0 \zeta(x,0,a)dxda + \int_{\Omega_T} \gamma(uw - u^2)\zeta(x,t,0)dxdt = 0,\end{aligned} \tag{5.7}$$

where $\theta(\cdot,\cdot)$ is defined by (2.14). This suggests that we give the following notion (compare with Definition 2.1).

DEFINITION 5.1. *Let (w,u,χ) be a weak solution of* (1.11)–(1.14) *on Ω_T and let ρ_0 be given satisfying* (5.4)–(5.5). *We call $\rho \in L^\infty(\Omega_T; \mathcal{M}(0,\infty))$ a weak infective age distribution associated with (w,u,χ) on $\Omega_T \times [0,\infty)$ if $\rho(x,t;\cdot)$ is a nonnegative measure for almost all (x,t), having the property*

$$\int_0^\infty \rho(x,t,da) = u(x,t) \quad a.e.\ (x,t) \in \Omega_T, \tag{5.8}$$

and for any $\zeta \in C^\infty(\bar{\Omega}_T^0 \times [0,\infty))$,

$$\begin{aligned}&\int_{\Omega_T \times [0,\infty)} \{-\zeta_t - \zeta_a + (\lambda+\beta)\zeta\}\rho(x,t,da)dxdt\\ &+ \int_{\Omega_T \times [0,\infty)} \frac{1}{2}\nabla w^2 \cdot \nabla \zeta \theta(w,\rho(x,t,da))dxdt\\ &+ \int_{\Omega \times [0,\infty)} \rho_0 \zeta(x,0,a)dxda + \int_{\Omega_T} \gamma(uw - \chi)\zeta(x,t,0)dxdt = 0,\end{aligned} \tag{5.9}$$

If the above holds for all $T > 0$, we simply call ρ a weak infective age distribution associated with (w,u,χ).

We note (see (2.14)) that the function θ is nonnegative and for any bounded open set $Q \subset \Omega_T$,

$$\int_{Q \times [0,\infty)} \theta(w(x,t),\rho(x,t,da))dxdt = \int_Q \frac{u(x,t)}{w(x,t)}dxdt \leq \operatorname{meas} Q.$$

Thus, the term involving $\theta(w, \rho(x,t,da))$ on the left-hand side of (5.9) makes sense.

Our main result of this section is the following theorem.

THEOREM 5.2. *Let ρ_0 satisfy* (5.4)–(5.5) *and (w_0, u_0) be as in previous sections. Then there exists a weak solution (w,u,χ) of* (1.11)–(1.14) *for which there is an associated weak infective age distribution ρ.*

To prove the above result we let $0 < \varepsilon, \delta \leq 1$ and let $(w_{\varepsilon,\delta}, u_{\varepsilon,\delta})$ be the solution of (3.1)–(3.4). Consider the following system:

$$\begin{cases} \rho_t + \rho_a = \nabla\rho \cdot \nabla w_{\varepsilon,\delta} + (\Delta w_{\varepsilon,\delta} - \lambda - \beta)\rho, \\ \qquad\qquad (x,t,a) \in \mathbb{R}^n \times \mathbb{R}^+ \times \mathbb{R}^+, \\ \rho\,|_{t=0} = \rho_0^\varepsilon \equiv \rho_0 + \varepsilon\gamma(w_0 - u_0), \qquad (x,a) \in \mathbb{R}^n \times \mathbb{R}^+, \\ \rho\,|_{a=0} = \gamma[u_{\varepsilon,\delta} w_{\varepsilon,\delta} - u_{\varepsilon,\delta}^2] \equiv \theta^{\varepsilon,\delta}, \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}^+. \end{cases} \tag{5.10}$$

Note that

$$\rho_0^\varepsilon(x,0) = \gamma(x,0)[u_0^\varepsilon(x) w_0^\varepsilon(x) - u_0^\varepsilon(x)^2] = \theta^{\varepsilon,\delta}(x,0) \quad \forall x \in \mathbb{R}^n. \tag{5.11}$$

This compatibility condition is needed to obtain a classical solution of (5.10).

We have the following lemma.

LEMMA 5.3. *There exists a unique classical solution $\rho_{\varepsilon,\delta}$ of* (5.10), *which is given by*

$$\rho_{\varepsilon,\delta}(\bar{x},\bar{t},\bar{a}) = \begin{cases} \rho_0^\varepsilon(\psi(0;\bar{x},\bar{t}),\bar{a}-\bar{t}) e^{\int_0^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt}, & \bar{t} \leq \bar{a}, \\ \theta^{\varepsilon,\delta}(\psi(\bar{t}-\bar{a};\bar{x},\bar{t}),\bar{t}-\bar{a}) e^{\int_{\bar{t}-\bar{a}}^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt}, & \bar{t} \geq \bar{a}, \end{cases} \tag{5.12}$$

where $\psi(\cdot;\bar{x},\bar{t}) \equiv \psi_{\varepsilon,\delta}(\cdot;\bar{x},\bar{t})$ is the characteristic associated with $w_{\varepsilon,\delta}$:

$$\begin{cases} \dfrac{d\psi}{dt} = -\nabla w_{\varepsilon,\delta}(\psi,t), \\ \psi\,|_{t=\bar{t}} = \bar{x}, \end{cases} \tag{5.13}$$

and $\ell = \Delta w_{\varepsilon,\delta} - \lambda - \beta$. Moreover, this solution satisfies

$$\int_0^\infty \rho_{\varepsilon,\delta}(x,t,a)da = u_{\varepsilon,\delta}(x,t), \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}^+. \tag{5.14}$$

Proof. Suppose ρ is a classical solution of (5.10). We let ψ be defined by (5.13) and set

$$\widetilde{\rho}(t) = \rho(\psi(t;\bar{x},\bar{t}),t,t-\bar{t}+\bar{a}), \tag{5.15}$$

Then we see that

$$\frac{d}{dt}\widetilde{\rho}(t) = \rho_t + \rho_a - \nabla\rho \cdot \nabla w_{\varepsilon,\delta} = \ell\widetilde{\rho}(t); \qquad \widetilde{\rho}(\bar{t}) = \rho(\bar{x},\bar{t},\bar{a}). \tag{5.16}$$

Hence we can obtain that ρ has the form (5.12). This gives the uniqueness of classical solutions of (5.10). Next we let $\rho_{\varepsilon,\delta}$ be given by (5.12). Then, as in the proof of

Lemma 3.7 we are able to show, with some lengthy but straightforward computation, that $\rho_{\varepsilon,\delta}$ is a classical solution of (5.10). Now we compute

$$
\begin{aligned}
\int_0^\infty \rho_{\varepsilon,\delta}(\bar{x},\bar{t},\bar{a})d\bar{a} &= \int_0^{\bar{t}} \theta^{\varepsilon,\delta}(\psi(\bar{t}-\bar{a};\bar{x},\bar{t}),\bar{t}-\bar{a})e^{\int_{\bar{t}-\bar{a}}^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt}d\bar{a} \\
&\qquad + e^{\int_0^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt}\int_{\bar{t}}^\infty \rho_0^\varepsilon(\psi(0;\bar{x},\bar{t}),\bar{a}-\bar{t})d\bar{a} \\
&= e^{\int_0^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt}\Big\{\int_0^{\bar{t}} \theta^{\varepsilon,\delta}(\psi(\bar{t}-\bar{a};\bar{x},\bar{t}),\bar{t}-\bar{a})e^{-\int_0^{\bar{t}-\bar{a}} \ell(\psi(t;\bar{x},\bar{t}),t)dt}d\bar{a} \\
&\qquad + \int_0^\infty \rho_0^\varepsilon(\psi(0;\bar{x},\bar{t}),\bar{a})d\bar{a}\Big\} \\
&= e^{\int_0^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt}u_0^\varepsilon(\psi(0;\bar{x},\bar{t}) \\
&\quad + \int_0^{\bar{t}} (\gamma u_{\varepsilon,\delta}w_{\varepsilon,\delta} - \gamma u_{\varepsilon,\delta})(\psi(\tau;\bar{x},\bar{t}),\tau)e^{-\int_\tau^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt}d\tau = u_{\varepsilon,\delta}(\bar{x},\bar{t}).
\end{aligned}
\tag{5.17}
$$

This completes the proof of the lemma. □

We note that (see (3.30))

$$
e^{\int_0^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt} = \frac{w_{\varepsilon,\delta}(\bar{x},\bar{t})}{w_0^\varepsilon(\psi(0;\bar{x},\bar{t}))}e^{\int_0^{\bar{t}}(c-\lambda-\beta)(\psi(t;\bar{x},\bar{t}),t)dt},
\tag{5.18}
$$

$$
e^{\int_{\bar{t}-\bar{a}}^{\bar{t}} \ell(\psi(t;\bar{x},\bar{t}),t)dt} = \frac{w_{\varepsilon,\delta}(\bar{x},\bar{t})}{w^{\varepsilon,\delta}(\psi(\bar{t}-\bar{a};\bar{x},\bar{t}),\bar{t}-\bar{a})}e^{\int_{\bar{t}-\bar{a}}^{\bar{t}}(c-\lambda-\beta)(\psi(t;\bar{x},\bar{t}),t)dt},
\tag{5.19}
$$

where (see (3.22))

$$
c-\lambda-\beta = -\left(\frac{\lambda E_\delta(w_{\varepsilon,\delta}-\varphi_\varepsilon * u_{\varepsilon,\delta})}{w_{\varepsilon,\delta}}+\beta\right) \le 0.
\tag{5.20}
$$

Thus, from (5.12) we obtain that for $\bar{t} \le \bar{a}$ (see (3.5)–(3.6)),

$$
0 \le \rho_{\varepsilon,\delta}(\bar{x},\bar{t},\bar{a}) \le \frac{\|\rho_0\|_{C^0(\Omega)}}{\varepsilon} + M(\|w_0\|_{C^0(\Omega)}+\varepsilon)^2
\tag{5.21}
$$

and for $\bar{t} \ge \bar{a}$,

$$
0 \le \rho_{\varepsilon,\delta}(\bar{x},\bar{t},\bar{a}) \le M(\|w_0\|_{C^0(\Omega)}+\varepsilon)^2.
\tag{5.22}
$$

This proves the following corollary.

COROLLARY 5.4. *The solution $\rho_{\varepsilon,\delta}$ satisfies*

$$
\begin{aligned}
0 \le &\rho_{\varepsilon,\delta}(\bar{x},\bar{t},\bar{a}) \le \frac{\|\rho_0\|_{C^0(\Omega)}}{\varepsilon} + M(\|w_0\|_{C^0(\Omega)}+\varepsilon)^2 \\
&\forall (\bar{x},\bar{t},\bar{a}) \in \mathbb{R}^n \times \mathbb{R}^+ \times \mathbb{R}^+.
\end{aligned}
\tag{5.23}
$$

Now we are ready to prove Theorem 5.2.

Proof of Theorem 5.2. First we will let $\delta \to 0$. By (4.5)–(4.6) we know that

$$\psi_{\varepsilon,\delta}(t;\bar{x},\bar{t}) \to \psi_\varepsilon(t;\bar{x},\bar{t}), \qquad \delta \to 0, \tag{5.24}$$

uniformly for $(t,\bar{x},\bar{t})$ in any compact set, where ψ_ε is the characteristic associated with w_ε. Thus, by (5.12) and (5.18)–(5.19),

$$\rho_{\varepsilon,\delta}(\bar{x},\bar{t},\bar{a}) \to \rho_\varepsilon(\bar{x},\bar{t},\bar{a}), \qquad \delta \to 0, \tag{5.25}$$

uniformly for all $(\bar{x},\bar{t},\bar{a})$ in compact sets and

$$\rho_\varepsilon(\bar{x},\bar{t},\bar{a}) = \begin{cases} \rho_0^\varepsilon(\psi_\varepsilon(0;\bar{x},\bar{t}),\bar{a}-\bar{t})e^{\int_0^{\bar{t}} \ell_\varepsilon(\psi_\varepsilon(t;\bar{x},\bar{t}),t)dt}, & \bar{t} \le \bar{a}, \\ \theta^\varepsilon(\psi_\varepsilon(\bar{t}-\bar{a};\bar{x},\bar{t}),\bar{t}-\bar{a})e^{\int_{\bar{t}-\bar{a}}^{\bar{t}} \ell_\varepsilon(\psi_\varepsilon(t;\bar{x},\bar{t}),t)dt}, & \bar{t} \ge \bar{a} \end{cases} \tag{5.26}$$

with $\theta^\varepsilon = \gamma(u_\varepsilon w_\varepsilon - u_\varepsilon^2)$ and $\ell_\varepsilon = \Delta w_\varepsilon - \lambda - \beta$. By (5.14) and (5.23), using the dominated convergence theorem, we have

$$\int_0^\infty \rho_\varepsilon(x,t,a)da = u_\varepsilon(x,t) \quad \forall (x,t) \in \Omega_T. \tag{5.27}$$

On the other hand, using the variational form for $\rho_{\varepsilon,\delta}$ and sending $\delta \to 0$, we have

$$\begin{aligned} &\int_{\Omega_T \times [0,\infty)} \Big\{ [-\zeta_t - \zeta_a + (\lambda+\beta)\zeta]\rho_\varepsilon + \frac{1}{2}\nabla w_\varepsilon^2 \cdot \nabla\zeta\theta(w_\varepsilon,\rho_\varepsilon) \Big\} dxdtda \\ &+ \int_{\Omega \times [0,\infty)} \rho_0 \zeta(x,0,a)dxda + \int_{\Omega_T} \gamma(u_\varepsilon w_\varepsilon - u_\varepsilon^2)\zeta(x,t,0)dxdt = 0. \end{aligned} \tag{5.28}$$

Next we need to send $\varepsilon \to 0$. By (5.27) and (5.23) we see that for any $A > 0$, $\{\rho_\varepsilon\}_{\varepsilon>0}$ is bounded in $L^\infty(\Omega_T; L^1(0,A))$. Since

$$L^\infty(\Omega_T; L^1(0,A)) \hookrightarrow L^2(\Omega_T; \mathcal{M}(0,A)) = L^2(\Omega_T; C([0,A]))^*, \tag{5.29}$$

we may let (using the diagonalization argument)

$$\rho_\varepsilon \to \rho \quad \text{weakly in } L^2(\Omega_T; \mathcal{M}(0,A)) \quad \forall A > 0. \tag{5.30}$$

Then, for any $\zeta \in C_0^\infty(\Omega_T \times [0,\infty))$, to obtain (5.9) we only need to prove the following:

$$\int_{\Omega_T \times [0,\infty)} \nabla w_\varepsilon^2 \cdot \nabla\zeta\theta(w_\varepsilon,\rho_\varepsilon)dxdtda \to \int_{\Omega_T \times [0,\infty)} \nabla w^2 \cdot \nabla\zeta\theta(w,\rho(x,t,da))dxdt. \tag{5.31}$$

It is not hard to see that $\{\theta(w_\varepsilon,\rho_\varepsilon)\}$ is bounded in $L^\infty(\Omega_T; \mathcal{M}(0,\infty))$. Thus we may assume that

$$\theta(w_\varepsilon,\rho_\varepsilon) \to \widetilde{\theta} \quad \text{weakly}^* \text{ in } L^\infty(\Omega_T; \mathcal{M}(0,\infty)). \tag{5.32}$$

By (5.30) and the convergence $w_\varepsilon \to w$ (uniformly in any compact sets; see §4), we can easily identify that $\widetilde{\theta} = \theta(w(x,t),\rho(x,t,da))$. On the other hand, from the proof of Theorem 4.6 we have $\nabla w_\varepsilon^2 \to \nabla w^2$ almost everywhere and weakly in $L^2(\Omega_T)$. Thus,

by Lemma 4.5 we can prove (5.31). Then, passing to the limit in (5.28), we obtain (5.9). Finally, by (5.27) and (5.30) we immediately obtain that for any $\xi \in C_0^\infty(\Omega_T)$,

$$\begin{aligned}\int_{\Omega_T} u\xi dxdt &= \lim_{\varepsilon\to 0}\int_{\Omega_T} u_\varepsilon \xi dxdt \\ &= \lim_{\varepsilon\to 0}\int_{\Omega_T\times[0,\infty)} \xi\rho_\varepsilon dxdtda = \int_{\Omega_T}\xi\left[\int_0^\infty \rho(x,t,da)\right]dxdt.\end{aligned} \tag{5.33}$$

Hence (5.7) follows. □

Remark 5.5. If there exists a $\sigma > 0$ such that

$$w_0(x) \geq \sigma \quad \forall x \in \mathbb{R}^n, \tag{5.34}$$

then (5.23) can be replaced by

$$\begin{aligned}0 \leq \rho_{\varepsilon,\delta}(\bar{x},\bar{t},\bar{a}) \leq \frac{\|\rho_0\|_{C^0(\Omega)}}{\sigma} + M(\|w_0\|_{C^0(\Omega)}+\varepsilon)^2 \\ \forall(\bar{x},\bar{t},\bar{a}) \in \mathbb{R}^n\times\mathbb{R}^+\times\mathbb{R}^+.\end{aligned} \tag{5.35}$$

In this case we see that the weak infective age distribution ρ is actually in $L^\infty(\mathbb{R}^n \times [0,\infty); L^1(0,\infty))$.

6. One-dimensional case. In this section we study the one-dimensional case. We show that if (w,u,χ) is a weak solution of (1.11)–(1.14), then we have $\chi = u^2$ near the points at which $w_0(x) > 0$. In particular, if for some $\eta > 0$, $w_0(x) \geq \eta$ for all $x \in \mathbb{R}$, then $\chi = u^2$ everywhere.

Let us start with the following lemma.

LEMMA 6.1. *Let* $(w_\varepsilon, u_\varepsilon)$ *be the solution of* (3.2)–(3.4) *and* (4.1). *Then, for any* $\eta > 0$ *with* $w_0(x) + \varepsilon \geq \eta$, *we have*

$$\|u_\varepsilon\|_{C^1(\bar{\Omega}_T)} \leq C(\eta,T)(1+\|\nabla w_\varepsilon\|_{C(\bar{\Omega}_T)})^2. \tag{6.1}$$

Proof. By Lemma 3.3, u_ε can be expressed by (3.11), where ψ is the characteristic (3.7) corresponding to w_ε, and $h = \Delta w_\varepsilon + \gamma w_\varepsilon - \lambda - \beta$. Since w_ε solves (4.1), we have

$$\begin{aligned}&\int_0^{\bar{t}} \Delta w_\varepsilon(\psi(t;\bar{x},\bar{t}),t)dt \\ &= \int_0^{\bar{t}}\left[\frac{1}{w_\varepsilon}\frac{d}{dt}w_\varepsilon(\psi(t;\bar{x},\bar{t}),t) + \lambda - \frac{\lambda E_0(w_\varepsilon - \varphi_\varepsilon * u_\varepsilon)}{w_\varepsilon}\right]dt \\ &= \log\left[\frac{w_\varepsilon(\bar{x},\bar{t})}{w_0^\varepsilon(\psi(0;\bar{x},\bar{t}))}\right] + \int_0^{\bar{t}}\left[\lambda - \lambda\frac{E_0(w_\varepsilon-\varphi_\varepsilon*u_\varepsilon)}{w_\varepsilon}\right]dt.\end{aligned} \tag{6.2}$$

On the other hand, $D_{\bar{x}}\psi(t;\bar{x},\bar{t})$ solves (3.9) (as a function of t). Hence, by computation analogous to (6.2) we get

$$D_{\bar{x}}\psi(t;\bar{x},\bar{t}) = e^{\int_t^{\bar{t}} w_{\varepsilon xx}(\psi(t;\bar{x},\bar{t}),t)dt} = \frac{w_\varepsilon(\bar{x},\bar{t})}{w_\varepsilon(\psi(t;\bar{x},\bar{t}),t)} e^{\int_0^{\bar{t}}[\lambda - \lambda\frac{E_0(w_\varepsilon-\varphi_\varepsilon*u_\varepsilon)}{w_\varepsilon}]dt}. \tag{6.3}$$

By the maximum principle and (4.1) we have

$$w_\varepsilon(x,t) \geq \eta e^{-MT}, \qquad (x,t) \in \Omega_T.$$

Thus, from (6.3) and (3.8) we obtain

$$|D_{\bar{x}}\psi(t;\bar{x},\bar{t})| \le C(\eta,T), \tag{6.4}$$

$$|D_{\bar{t}}\psi(t;\bar{x},\bar{t})| \le C(\eta,T)\|\nabla w_\varepsilon\|_{C(\bar{\Omega}_T)}. \tag{6.5}$$

By directly taking derivatives in (3.11) we can get (6.1). □

THEOREM 6.2. *Let $\eta > 0$ and*

$$w_0(x) \ge \eta \quad \forall x \in \mathbb{R}. \tag{6.6}$$

Let (w,u,χ) be a weak solution of (1.11)–(1.14). *Then*

$$\chi(x,t) = u(x,t)^2 \quad \forall (x,t) \in \mathbb{R}\times(0,\infty). \tag{6.7}$$

Proof. Because of (6.6), by the Schauder-type estimates we have that for any $T > 0$, there exists a constant $C(T)$ such that

$$\|w_\varepsilon\|_{C^{2,\alpha}(\bar{\Omega}_T)} \le C(T)(1 + \|u_\varepsilon\|_{C^\alpha(\bar{\Omega}_T)}). \tag{6.8}$$

It follows from (6.1) that both $\|w_\varepsilon\|_{C^{2,\alpha}(\bar{\Omega}_T)}$ and $\|u_\varepsilon\|_{C^1(\bar{\Omega}_T)}$ are uniformly bounded. Since (u,χ) is the weak limit of $(u_\varepsilon, u_\varepsilon^2)$, we obtain (6.7). □

Next we would like to consider the case in which w_0 does not have the uniform positive lower bound. In this case we have the following local result.

THEOREM 6.3. *Let $x_0 \in \mathbb{R}$ be such that $w(x_0) > 0$. Then, there exists a neighborhood $B_1(x_0)$ of x_0 in $\mathbb{R}$ and a $T_1 > 0$ such that*

$$\chi(x,t) = u(x,t)^2 \qquad \forall (x,t) \in B_1(x_0)\times[0,T_1]. \tag{6.9}$$

Proof. From Lemma 4.4 we know that there exists a neighborhood $B(x_0)$ of x_0 and a $T > 0$ such that for some $\eta > 0$,

$$w(x,t) \ge \eta, \qquad (x,t) \in B(x_0)\times[0,T]. \tag{6.10}$$

Then it follows that for $\varepsilon > 0$ small enough,

$$w_\varepsilon(x,t) \ge \frac{1}{2}\eta, \qquad (x,t) \in B(x_0)\times[0,T]. \tag{6.11}$$

Hence, by the interior $C^{1+\alpha}$-estimates we have some constant $C > 0$ independent of $\varepsilon > 0$, such that

$$\|w_\varepsilon\|_{C^{1,\alpha}(\bar{B}(x_0)\times[0,T])} \le C \quad \forall \varepsilon > 0. \tag{6.12}$$

Let ψ_ε be the characteristics corresponding to w_ε. By (3.7) and (6.12) we can find a small neighborhood $B_1(x_0)\times[0,T_1] \subset B(x_0)\times[0,T]$ such that for any $(\bar{x},\bar{t}) \in B_1(x_0)\times[0,T_1]$, $\psi_\varepsilon(t;\bar{x},\bar{t})$ always stays in $B(x_0)$ for $t \in [0,T_1]$. Therefore we have

$$|D_{\bar{x}}\psi_\varepsilon(t;\bar{x},\bar{t})| \le C(\eta,T_1), \tag{6.13}$$

$$|D_{\bar{t}}\psi_\varepsilon(t;\bar{x},\bar{t})| \le C(\eta,T_1)\|\nabla w_\varepsilon\|_{C(\bar{B}(x_0)\times[0,T_1])} \tag{6.14}$$

for all $t \in [0, t_1]$ and $(\bar{x}, \bar{t}) \in B_1(x_0) \times [0, T_1]$. By the same argument as the one in the proof of Theorem 6.2, we obtain (6.9). □

COROLLARY 6.4. *Let $I \subset \mathbb{R}$ be a compact set such that*

$$w_0(x) > 0 \quad \forall x \in I. \tag{6.15}$$

Then there exists a neighborhood $\mathcal{O} \subset \mathbb{R} \times [0, \infty)$ of $I \times \{0\}$ such that

$$\chi(x,t) = u(x,t)^2, \qquad (x,t) \in \mathcal{O}. \tag{6.16}$$

The proof is immediate.

REFERENCES

[1] M. BERTSCH, M. E. GURTIN, D. HILHORST, AND L. A. PELETER, *On interacting populations that disperse to avoid crowding: Preservation of segregation*, J. Math. Biol., 23 (1985), pp. 1–13.

[2] S. N. BUSENBERG AND K. P. HADELER, *Demography and epidemics*, Math. Biosci., 101 (1990), pp. 63–74.

[3] S. BUSENBERG AND M. IANNELLI, *A class of nonlinear diffusion problems in age-dependent population*, Nonlinear Anal., 7 (1981), pp. 501–529.

[4] ———, *A degenerate nonlinear diffusion problem in age-structured population dynamics*, Nonlinear Anal., 7 (1981), pp. 1411–1429.

[5] E. DIBENEDETTO AND A. FRIEDMAN, *Regularity of solutions of nonlinear degenerate parabolic systems*, J. Reine Angew Math., 349 (1984), pp. 83–128.

[6] ———, *Hölder estimates for nonlinear degenerate parabolic systems*, J. Reine Angew Math., 357 (1985), pp. 1–22.

[7] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Inc., Englewood Cliffs, NJ, 1964.

[8] M. E. GURTIN, *A system of equations for age-dependent population diffusion*, J. Theory Biol., 40 (1973), pp. 389–392.

[9] M. E. GURTIN AND R. C. MACCAMY, *On the diffusion of biological populations*, Math. Biosci., 33 (1977), pp. 35–49.

[10] C. HUANG, *An age-dependent population model with nonlinear diffusion in $\mathbb{R}^n$*, Quart. Appl. Math., to appear.

[11] H. INABA, *Threshold and stability results for an age-structured epidemic model*, J. Math. Biol., 28 (1990), pp. 411–434.

[12] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[13] M. LANGLAIS, *A nonlinear problem in age dependent population diffusion*, SIAM J. Math. Anal., 16 (1985), pp. 510–529.

[14] R. C. MACCAMY, *A population model with nonlinear diffusion*, J. Differential Equations, 39 (1981), pp. 52–72.

[15] J. D. MURRAY, *Mathematical Biology*, Springer-Verlag, New York, 1989.

[16] L. TARTAR, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot–Watt Symposium, Pitman, London, 1979, pp. 136–212.

[17] E. VENTURINO, *The influence of diseases on Lotka–Volterra systems*, Institute for Mathematics and Its Applications, 913, preprint.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1611-1643, November 1995

013

SINGULAR PERTURBATION THEORY FOR HOMOCLINIC ORBITS IN A CLASS OF NEAR-INTEGRABLE DISSIPATIVE SYSTEMS*

GREGOR KOVAČIČ†

Abstract. This paper presents a new unified theory of orbits homoclinic to resonance bands in a class of near-integrable dissipative systems. It describes three sets of conditions, each of which implies the existence of homoclinic or heteroclinic orbits that connect equilibria or periodic orbits in a resonance band. These homoclinic and heteroclinic orbits are born under a given small dissipative perturbation out of a family of heteroclinic orbits that connect pairs of points on a circle of equilibria in the phase space of the nearby integrable system. The result is a constructive method that may be used to ascertain the existence of orbits homoclinic to objects in a resonance band, as well as to determine their precise shape, asymptotic behavior, and bifurcations in a given example. The method is a combination of the Melnikov method and geometric singular perturbation theory for ordinary differential equations.

Key words. Melnikov method, geometric singular perturbation theory, inner and outer limits, homoclinic orbits, resonance

AMS subject classifications. 34A26, 34A47, 34C35, 34C37, 34D15

1. Introduction.

Completely integrable Hamiltonian systems are a fairly rare occurrence. Nevertheless, since they can be solved explicitly, they are the first step in the description of many fundamental physical phenomena, such as the motion of rigid bodies or the circling of the earth around the sun. Knowledge of the phase-space properties of these special idealized systems is used in conjunction with perturbation theory to describe more realistic physical problems, for instance, problems that exhibit irregular or chaotic behavior.

The perturbation method most commonly used to show the presence of chaotic dynamics in near-integrable systems is the Melnikov method. First developed for time-periodically perturbed planar systems [1]–[5], it was soon generalized to cover multi-degree-of-freedom systems as well [6]–[17]. This method is particularly convenient for multidimensional Hamiltonian systems, where it combines with the Kolmogorov–Arnold–Moser theory [18]–[22] to yield the existence of Smale-horseshoe chaos and Arnold diffusion in many problems [7]–[9], [23]–[29].

The use of the multidimensional Melnikov method for near-integrable dissipative systems is restricted to special cases [13], [14], [16], [30], [31]. In all of these cases, averaging or some singular perturbation method must be used together with the Melnikov method to show the existence of some homoclinic or heteroclinic orbits, whose presence causes nearby phase points to behave chaotically. The singular perturbation aspect is most stressed in [30]. That paper shows how to construct a spiral-saddle connection out of a circle of equilibria and a two-dimensional surface of heteroclinic orbits connecting certain pairs of points on this circle. The spiral-saddle itself, as well as the spiraling part of the connection, are born out of the circle of equilibria under perturbation; hence the singular nature of the problem.

This paper presents a geometric theory of phenomena that can emerge under perturbation out of a manifold of orbits homoclinic to a circle of equilibria, which lies on an unstable invariant annulus in the phase space of an $(n+1)$-degree-of-freedom

* Received by the editors March 8, 1993; accepted for publication (in revised form) March 3, 1994. This research was supported by Department of Energy grant DE–FG02–93ER25154 and National Science Foundation grant DMS-9403750.

† Mathematical Sciences Department, Rensselaer Polytechnic Institute, Troy, New York 12180.

integrable dynamical system. In particular, the paper discusses a number of possible homoclinic and heteroclinic orbits connecting equilibria and periodic orbits that lie inside a resonance band [32]–[34] created by the perturbation out of the circle of equilibria.

The main results of this paper are presented in the three theorems in §3. These theorems describe the various geometric situations that give rise to homoclinic or heteroclinic orbits connecting objects in a resonance band. Theorem 1 describes orbits that result from intersections of $(n+1)$-dimensional stable and $(n+1)$-dimensional unstable manifolds; Theorem 2 describes orbits that result from intersections of $(n+2)$-dimensional stable and $(n+1)$-dimensional unstable manifolds; and Theorem 3 describes orbits that result from intersections of $(n+2)$-dimensional stable and n-dimensional unstable manifolds.

The result presented in [30] is a special example of this paper's Theorem 3. A very similar example is given in [31]. An example of Theorem 1 is presented in [35]. This example shows how to construct orbits homoclinic to saddles in a resonance band that have purely real eigenvalues. A similar example is studied in [36], in which an independent proof is given for a special case of this paper's Theorem 1, and careful numerical calculations are performed that support the theoretical findings.

The proofs of the three main theorems of this paper that are described in §7 require a fair amount of background, which is outlined in §§4–6. In particular, all three proofs follow virtually the same geometric idea, which consists of three main steps. The first step is outlined in §4. In this step, an unperturbed unstable invariant annulus, which contains a circle of equilibria and is connected to itself by an $(n+2)$-dimensional homoclinic manifold, is shown to persist under perturbation, and the Melnikov method is used to ascertain whether a two-dimensional homoclinic manifold of orbits that are biasymptotic to this persisting annulus exists. In the second step, rescaling is used to describe the dynamics in the resonance band that is created by the perturbation on the persisting annulus out of the unperturbed circle of equilibria. This step is developed in §5. In the third step, geometric singular perturbation theory [37], [38] is used to connect the homoclinic dynamics, which are transverse to the persisting annulus, to the dynamics along this annulus in order to describe the precise asymptotic behavior of the homoclinic orbits. This step is carried out in §§6 and 7.

A Hamiltonian counterpart of this paper was developed in [39]–[41]. For two-degree-of-freedom systems, the Hamiltonian result is very general, because the Melnikov function used in that situation can be computed explicitly as an energy difference. The details of the Hamiltonian case have much in common with Theorem 1 of this paper. In particular, two crucial stepping stones in its proof are Propositions 7.1 and 7.2. However, the geometry of the Hamiltonian case is very different from the geometry described in the present paper. Namely, in the Hamiltonian case, the result describes two-dimensional surfaces of orbits homoclinic to nested families of periodic orbits as opposed to isolated homoclinic or heteroclinic orbits described in this paper. Moreover, orbits homoclinic to objects inside a resonance band are generic in the Hamiltonian case, but only occur on lower-dimensional submanifolds of the parameter space in most subcases of the dissipative case presented here.

Methods for finding orbits homoclinic to resonance bands may be applied to some of the systems that have undergone a change of variables into a frame rotating with the same frequency as an external force and subsequent averaging. Examples of such systems are [42] in the theory of Josephson's junctions; [25] in nonlinear fiber optics; [26], [28], [29] in laser–matter interaction; [13], [23], [24], [31], [43] in the theory

of water waves; and [27] in the theory of vibrating plates. The advantage of the methods presented in this paper's Theorems 1–3, as well as in the main theorem of [40], is that their hypotheses are easily verified in specific situations. In fact, their verification requires only algebraic manipulations. This makes the method described in this paper a potentially powerful tool for solving physical and engineering problems.

This paper is organized as follows. In §2, the problem of orbits homoclinic to resonance bands is set up. In §3, the main results of the paper are stated. In §4, results from persistence theory of normally hyperbolic invariant manifolds that are needed for the understanding of this paper are discussed, and a brief review of the multi-dimensional Melnikov method is given. In §5, an approach to analyzing resonance bands is explained. In §6, geometric singular perturbation theory is used to calculate local stable and unstable manifolds of objects in a resonance band. In §7, the three main theorems of this paper are proven. Finally, in §8, a simple example is shown to satisfy the conditions of the three main theorems for certain parameter values.

2. The setup. We consider systems of the form

$$\dot{x} = JD_xH(x,I) + \varepsilon g^x(x,I,\theta,\lambda), \tag{2.1a}$$

$$\dot{I} = \varepsilon g^I(x,I,\theta,\lambda), \tag{2.1b}$$

$$\dot{\theta} = \Omega(x,I) + \varepsilon g^\theta(x,I,\theta,\lambda). \tag{2.1c}$$

Here $x = (x_1,\dots,x_{2n}) \in \mathbb{R}^{2n}$, $I \in \mathbb{R}$, and $\theta \in S^1$; D_x denotes the partial derivatives with respect to x; $\lambda \in \mathbb{R}$ is a real parameter; $\varepsilon \ll 1$ is a small parameter; and

$$J = \begin{pmatrix} 0 & -Id \\ Id & 0 \end{pmatrix},$$

with Id being the $n \times n$ identity matrix.

When we set $\varepsilon = 0$, we obtain the unperturbed system

$$\dot{x} = JD_xH(x,I), \tag{2.2a}$$

$$\dot{I} = 0, \tag{2.2b}$$

$$\dot{\theta} = \Omega(x,I). \tag{2.2c}$$

We immediately note that equation (2.2a) is a one-parameter family of Hamiltonian systems for the variable x and can be analyzed independently of θ. Equation (2.2c) can be solved by quadrature, once equation (2.2a) has been solved.

We first make two assumptions about the system (2.2). The first assumption concerns its solvability.

ASSUMPTION 1. *For all I with $I_1 < I < I_2$, for some I_1 and I_2, the system (2.2a) is completely integrable; that is, there exists a smooth family of n integrals of motion, $K_1(x,I) = H(x,I)$, $K_2(x,I),\dots,K_n(x,I)$, whose gradients $D_xK_1(x,I)$, $D_xK_2(x,I),\dots,D_xK_n(x,I)$ are linearly independent at all points x which are not equilibria of (2.2a) and pairwise satisfy the relationship*

$$\langle JD_xK_i(x,I), D_xK_j(x,I)\rangle = 0, \tag{2.3}$$

for all $i, j = 1, \ldots, n$.

This assumption implies that, at least in principle, solutions to equation (2.2a) may be obtained by quadratures; see, for instance, [17].

The second assumption introduces homoclinic orbits into the phase space of equations (2.2a).

ASSUMPTION 2. *For every* I *with* $I_1 < I < I_2$, *equation* (2.2a) *possesses a hyperbolic equilibrium* $x = X(I)$, *which varies smoothly with* I *and a manifold* $W(X(I))$ *of homoclinic orbits, connecting the equilibrium at* $x = X(I)$ *to itself.*

We remark that the stable and unstable manifolds $W^{\mathrm{s}}(X(I))$ and $W^{\mathrm{u}}(X(I))$ of the equilibrium $X(I)$ must both be n-dimensional, since the eigenvalues of the matrix $JD_x^2 H(X(I), I)$ come in pairs κ, $-\kappa$. The homoclinic manifold $W(X(I))$ must also be n-dimensional because of the linear independence of the gradients $D_x K_1(x, I), \ldots, D_x K_n(x, I)$. (See, for instance, [17, Prop. 4.1.3.])

Since the system (2.2a) is autonomous, all the solutions on the homoclinic manifold $W(X(I))$ can be represented in the form $x^h(t - t_0, I, \phi)$, where $\phi \in \mathbb{R}^{n-1}$ is a vector of parameters. A consistent parametrization of individual orbits in the manifold $W(X(I))$ can be obtained by setting $t_0 = 0$ and varying t.

In the full $(2n+2)$-dimensional phase space of the system (2.2), the family of equilibria at $x = X(I)$ forms a two-dimensional invariant annulus $\mathcal{M}$ foliated by periodic orbits with coordinates $x = X(I)$, I, and $\theta = \Omega(X(I), I)t + \theta_0$, with $I_1 < I < I_2$. The annulus $\mathcal{M}$ possesses $(n+2)$-dimensional stable and unstable manifolds $W^{\mathrm{s}}(\mathcal{M})$ and $W^{\mathrm{u}}(\mathcal{M})$, which are the unions over the interval $I_1 < I < I_2$ of the Cartesian products of the manifolds $W^{\mathrm{s}}(X(I))$ and $W^{\mathrm{u}}(X(I))$ with the angle θ, respectively. The manifolds $W^{\mathrm{s}}(\mathcal{M})$ and $W^{\mathrm{u}}(\mathcal{M})$ intersect along the $(n+2)$-dimensional homoclinic manifold $W(\mathcal{M})$, which is the union over the interval $I_1 < I < I_2$ of the Cartesian products of the homoclinic manifolds $W(X(I))$ with the angle θ, shown in Fig. 1. We remark that the homoclinic manifold $W(\mathcal{M})$ can be parametrized by t, I, ϕ, and θ_0 in the representation

$$x = x^h(t, I, \phi), \qquad I = I, \qquad \theta = \theta^h(t, I, \phi) + \theta_0, \tag{2.4}$$

with

$$\theta^h(t, I, \phi) = \int_0^t \Omega(x^h(s, I, \phi), I) ds.$$

It can also be represented implicitly by the set of equations

$$K_i(x, I) - K_i(X(I), I) = 0, \quad i = 1, \ldots, n, \tag{2.5}$$

which hold on the annulus $\mathcal{M}$ at $x = X(I)$ and, therefore, also on the homoclinic manifold $W(\mathcal{M})$.

To study orbits homoclinic to resonance bands, we make the following assumption.

ASSUMPTION 3. *For some* I_0 *with* $I_1 < I_0 < I_2$ *we have*

$$\Omega(X(I_0), I_0) = 0$$

with

$$\frac{d\Omega(X(I_0), I_0)}{dI} \neq 0.$$

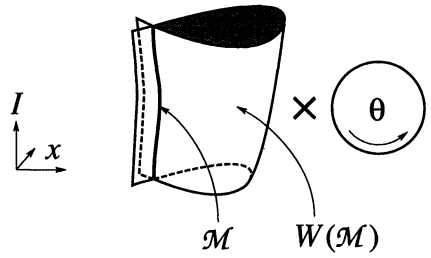


FIG. 1. *The invariant annulus $\mathcal{M}$ and its three-dimensional homoclinic manifold $W(\mathcal{M})$ are the Cartesian product of a circle with a curve segment filled with equilibria and its two-dimensional homoclinic manifold.*

This assumption implies that for $I = I_0$, the frequency of the periodic orbit on the annulus $\mathcal{M}$ passes through a simple zero, so that this periodic orbit is really a circle of equilibria. As will be shown in §5, this circle will break up under the given perturbation into a *resonance band*, which is the main object of this study.

Any equilibrium p on the circle is determined by its value of the angle $\theta = \theta(p)$. The unstable manifold of the point p is the set parametrized by the variables t and ϕ in the formulas $x = x^h(t, I_0, \phi)$, $I = I_0$, $\theta = \theta^h(t, I_0, \phi) + \theta_0(p, \phi)$ with the phase angle $\theta_0(p, \phi)$ defined by $\theta_0(p, \phi) = \theta(p) - \theta^h(-\infty, I_0, \phi)$. This set is an n-dimensional manifold, in general foliated by heteroclinic connections between p and other equilibria on the circle, whose θ coordinates are given by the formula $\theta = \theta(p) + \Delta\theta(\phi)$, in which the expression $\Delta\theta(\phi) = \theta^h(\infty, I_0, \phi) - \theta^h(-\infty, I_0, \phi)$ depends only on the value of the parameter vector ϕ and not on the initial equilibrium p, because of the phase symmetry of equations (2.2) in θ. Similar statements hold for the stable manifold $W^s(p)$.

As mentioned above, the circle of equilibria at $I = I_0$ breaks up under the perturbation into a resonance band. For this resonance band to contain only a finite number of discrete equilibria, we assume the following.

ASSUMPTION 4. *At any fixed value of the parameter λ, the function $g^I(X(I_0), I_0, \theta, \lambda)$ has only finitely many simple zeros in θ for $0 \le \theta \le 2\pi$.*

Finally, we define the Melnikov vector, $\mathbf{M}(I, \phi, \theta_0, \lambda)$, whose n components $M_i(I, \phi, \theta_0, \lambda)$ are given by the formulas

$$M_i(I, \phi, \theta_0, \lambda) = \int_{-\infty}^{\infty} \langle \mathbf{n}_i, g \rangle \, dt, \tag{2.6}$$

where

$$\begin{aligned} \mathbf{n}_i &= \left(D_x K_i(x^h(t, I, \phi), I), D_I K_i(x^h(t, I, \phi), I) - \frac{dK_i}{dI}(X(I), I), 0 \right) \\ &= \left(D_x K_i(x^h(t, I, \phi), I), D_I K_i(x^h(t, I, \phi), I) - D_I K_i(X(I), I), 0 \right), \end{aligned}$$

for $i = 1, \ldots, n$, are the n normals to the homoclinic manifold $W(\mathcal{M})$ that can be calculated from the equation (2.5), and $g = (g^x, g^I, g^\theta)$ is the $\mathcal{O}(\varepsilon)$ perturbation part of the vector field (2.1), calculated at $x = x^h(t, I, \phi)$, I, and $\theta = \theta^h(t, I, \phi) + \theta_0$; see [1]–[17]. The above two expressions for the normal $\mathbf{n}_i$ are equivalent because $D_x K_i(X(I), I) = 0$. (This follows from differentiating equation (2.3) with $j = 1$ upon x, substituting $x = X(I)$, and remembering that $x = X(I)$ is a hyperbolic equilibrium of the equation (2.2a), so that the matrix $JD^2H(X(I), I)$ is invertible; see [10] or [17, p. 407].)

For the rest of this paper, we assume the following.

ASSUMPTION 5. *For $I = I_0$ and some $\phi = \bar{\phi}$, $\theta_0 = \bar{\theta}_0$, and $\lambda = \bar{\lambda}$ the following two statements are true:*

1. $\mathbf{M}(I_0, \bar{\phi}, \bar{\theta}_0, \bar{\lambda}) = 0$,
2. $D_{(\phi,\theta_0)}\mathbf{M}(I_0, \bar{\phi}, \bar{\theta}_0, \bar{\lambda})$ *has maximal rank.*

3. The main results. In this section, we state the main results of this paper. They are described in Theorems 1–3, and follow from a series of preliminary results which we outline next. All the proofs and further details are relegated to §§4–7.

First, we observe that the results of Fenichel [44]–[46] imply that the annulus $\mathcal{M}$ and its stable and unstable manifolds $W^s(\mathcal{M})$ and $W^u(\mathcal{M})$ persist under perturbation as a locally invariant annulus $\mathcal{M}_\varepsilon$ and its stable and unstable manifolds $W^s(\mathcal{M}_\varepsilon)$ and $W^u(\mathcal{M}_\varepsilon)$. The precise nature of these manifolds will be discussed in Proposition 4.1. What is important for this outline is that the perturbed annulus $\mathcal{M}_\varepsilon$ can be written as a graph over the I and θ variables in the form

$$x = X_\varepsilon(I, \theta, \lambda, \varepsilon), \tag{3.1}$$

for some smooth function $X_\varepsilon(I, \theta, \lambda, \varepsilon)$ with $X_0(I, \theta, \lambda, 0) = X(I)$.

The circle of equilibria that exists on the unperturbed annulus $\mathcal{M}$ breaks up under perturbation into a *resonance band* lying on the perturbed annulus $\mathcal{M}_\varepsilon$. This resonance band is best described in the following way. We restrict the dynamics of equations (2.1) to the annulus $\mathcal{M}_\varepsilon$ using formula (3.1). Following [32]–[34], we then "blow up" the region near $I = I_0$ using the transformation $I = I_0 + \sqrt{\varepsilon}\, h$, rescale time using $\tau = \sqrt{\varepsilon} t$, and Taylor expand in $\sqrt{\varepsilon}$, to obtain the equations

$$h' = g^I(X(I_0), I_0, \theta, \lambda) + \mathcal{O}(\sqrt{\varepsilon}), \tag{3.2a}$$

$$\theta' = \frac{d\Omega}{dI}(X(I_0), I_0)\ h + \mathcal{O}(\sqrt{\varepsilon}), \tag{3.2b}$$

with $' = \frac{d}{d\tau}$. Higher-order terms in these equations can be easily computed just in terms of differentiations and algebraic operations alone, as implied by Proposition 5.

In the limit as $\varepsilon \to 0$, we obtain the *rescaled* or *outer system*

$$h' = g^I(X(I_0), I_0, \theta, \lambda), \tag{3.3a}$$

$$\theta' = \frac{d\Omega}{dI}(X(I_0), I_0)\ h. \tag{3.3b}$$

The outer system (3.3) can be derived from the *rescaled Hamiltonian*

$$\mathcal{H}(h, \theta, \lambda) = \frac{1}{2}\frac{d\Omega}{dI}(X(I_0), I_0)\, h^2 + V(\theta, \lambda), \tag{3.4}$$

with

$$V(\theta, \lambda) = -\int_0^\theta g^I(X(I_0), I_0, s, \lambda) ds,$$

via the canonical formulas

$$h' = -D_\theta \mathcal{H}(h, \theta, \lambda), \qquad \theta' = D_h \mathcal{H}(h, \theta, \lambda).$$

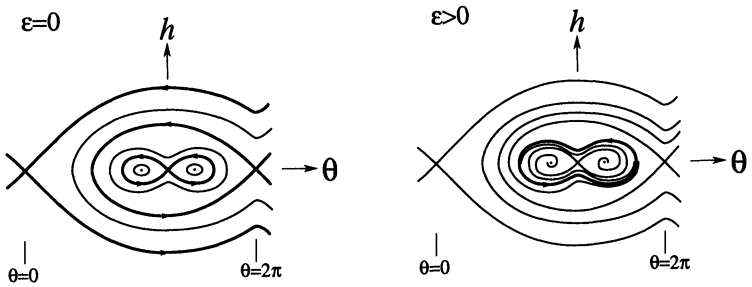


FIG. 2. *A typical phase portrait of a rescaled or outer system for* $\varepsilon = 0$ *and* $\varepsilon > 0$. *All the points whose* θ *coordinates differ by a multiple of* 2π *must be identified.*

Note that the limiting outer system (3.3) is Hamiltonian also when systems (3.2), (2.1), and even (2.2) are not.

System (3.2) can be investigated with the help of system (3.3) by a mixture of phase-plane and perturbation techniques. One approach to this investigation is outlined in §5. The phase portraits of a typical outer system (3.3) and its perturbed counterpart (3.2) are shown in Fig. 2.

To investigate orbits homoclinic or heteroclinic to possible equilibria and periodic orbits of equations (3.2) in the full $x - I - \theta$ phase space, we simply set $I = I_0 + \sqrt{\varepsilon}h$ in equations (2.1) and let $\varepsilon \to 0$. The resulting system is the *inner system*

$$\dot{x} = JD_xH(x, I_0)\,, \tag{3.5a}$$

$$\dot{h} = 0\,, \tag{3.5b}$$

$$\dot{\theta} = \Omega\,(x, I_0)\,. \tag{3.5c}$$

This system is a singular limit of equations (2.1) in the sense that the circle of equilibria for the unperturbed equations (2.2) at $I = I_0$, $x = X(I_0)$, $0 \le \theta \le 2\pi$ has been blown up into a cylinder of equilibria with $x = X(I_0)$, $0 \le \theta \le 2\pi$, and arbitrary h. Equilibria on this cylinder are unstable, and the structure of each circle $h =$ constant and the heteroclinic orbits that connect pairs of points on it is the same (including the solutions on the heteroclinic orbits) as the structure of the circle at $I = I_0$ and its homoclinic manifold $W(X(I_0))$. Thus, the $h - \theta$ cylinder of equilibria at $x = X(I_0)$ is connected to itself by an $(n+2)$-dimensional homoclinic manifold.

In what is to follow, we confine the values of h to the interval $-C < h < C$ with some large enough constant C. The annular portion of the $h - \theta$ cylinder between the circles $h = -C$ and $h = C$ will be called $\hat{\mathcal{M}}$. The annulus $\hat{\mathcal{M}}$ and its stable and unstable manifolds $W^{\mathrm{s}}(\hat{\mathcal{M}})$ and $W^{\mathrm{u}}(\hat{\mathcal{M}})$ are the limits as $\varepsilon \to 0$ of a nearby annulus $\hat{\mathcal{M}}_\varepsilon$ and its stable and unstable manifolds, $W^{\mathrm{s}}(\hat{\mathcal{M}}_\varepsilon)$ and $W^{\mathrm{u}}(\hat{\mathcal{M}}_\varepsilon)$. In the original $x - I - \theta$ coordinates, the annulus $\hat{\mathcal{M}}_\varepsilon$ and its stable and unstable manifolds $W^{\mathrm{s}}(\hat{\mathcal{M}}_\varepsilon)$ and $W^{\mathrm{u}}(\hat{\mathcal{M}}_\varepsilon)$ are just small pieces of the perturbed annulus $\mathcal{M}_\varepsilon$ and the manifolds $W^{\mathrm{s}}(\mathcal{M}_\varepsilon)$ and $W^{\mathrm{u}}(\mathcal{M}_\varepsilon)$, which shrink to zero like $\sqrt{\varepsilon}$ as $\varepsilon \to 0$.

The inner and outer systems are complementary in the following way. The inner system describes the structure of homoclinic orbits away from the annulus $\hat{\mathcal{M}}$, but the annulus $\hat{\mathcal{M}}$ itself consists of equilibria, and all the nontrivial dynamics on the nearby annulus $\hat{\mathcal{M}}_\varepsilon$ are lost in the inner limit. On the other hand, the $h - \theta$ cylinder

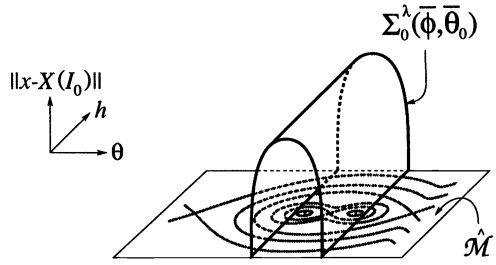


FIG. 3. *The limiting homoclinic manifold* $\Sigma_0^\lambda(\bar\phi,\bar\theta_0)$ *connects equilibria that lie on the line* $\theta=\bar\theta_0-\Delta\theta_-(\bar\phi)$ *to those that lie on the line* $\theta=\bar\theta_0+\Delta\theta_+(\bar\phi)$ *on the annulus* $\hat{\mathcal{M}}$. *Gray curves on* $\hat{\mathcal{M}}$ *represent the orbit structure on this annulus under the rescaled or outer system.*

$\hat{\mathcal{M}}$ for the outer equations (3.3) possesses nontrivial dynamics due to the rescaling of time; however, all the dynamics away from $x=X(I_0)$ are lost in this system. This situation is typical of singular perturbation problems in which we combine the information obtained from systems (3.5) and (3.3) to obtain useful information about the behavior of system (2.1) near the resonance band on the perturbed annulus $\mathcal{M}_\varepsilon$ at $I=I_0$.

We mentioned above that the annulus $\hat{\mathcal{M}}$ in the inner limit possesses an $(n+2)$-dimensional homoclinic manifold. We will show in Proposition 4.3 that a two-dimensional intersection surface, $\Sigma_\varepsilon^\lambda(\bar\phi,\bar\theta_0)$, of the manifolds $W^s(\hat{\mathcal{M}}_\varepsilon)$ and $W^u(\hat{\mathcal{M}}_\varepsilon)$ survives from this homoclinic manifold for nonzero ε. This surface corresponds to the transverse zero of the Melnikov vector at $I=I_0$, $\phi=\bar\phi$, $\theta_0=\bar\theta_0$, and $\lambda=\bar\lambda$, whose existence was assumed in Assumption 5. The surface $\Sigma_\varepsilon^\lambda(\bar\phi,\bar\theta_0)$ exists for all λ close enough to $\lambda=\bar\lambda$. In the inner limit, the surface $\Sigma_\varepsilon^\lambda(\bar\phi,\bar\theta_0)$ tends to the *limiting homoclinic intersection surface* $\Sigma_0^\lambda(\bar\phi,\bar\theta_0)$, shown in Fig. 3. This surface consists of those heteroclinic orbits connecting equilibria on the cylinder $\hat{\mathcal{M}}$ whose ϕ and θ_0 parameters equal $\phi=\bar\phi(I_0,\lambda)$, $\theta_0=\bar\theta_0(I_0,\lambda)$, where $\bar\phi(I,\lambda)$ and $\bar\theta_0(I,\lambda)$ are two smooth functions, defined for I and λ near $I=I_0$ and $\lambda=\bar\lambda$, with $\bar\phi(I_0,\bar\lambda)=\bar\phi$ and $\bar\theta_0(I_0,\bar\lambda)=\bar\theta_0$, that identically satisfy the equation $\mathbf{M}(I,\bar\phi(I,\lambda),\bar\theta_0(I,\lambda),\lambda)=0$. The heterodonic orbits on the limiting intersection surface $\Sigma_0^\lambda(\bar\phi,\bar\theta_0)$ are thus explicitly given by the formulas $x=x^h(t,I_0,\bar\phi(I_0,\lambda))$, $h=h$, $\theta=\theta^h(t,I_0,\bar\phi(I_0,\lambda))+\bar\theta_0(I_0,\lambda)$. These orbits thus emerge from the $h-\theta$ cylinder $\hat{\mathcal{M}}$ along the line $\theta=\bar\theta_0(I_0,\lambda)-\Delta\theta_-(\bar\phi(I_0,\lambda))$ and return to $\hat{\mathcal{M}}$ along the line $\theta=\bar\theta_0(I_0,\lambda)+\Delta\theta_+(\bar\phi(I_0,\lambda))$, where

$$(3.6)\quad \Delta\theta_+(\phi)=\int_0^\infty \Omega(x^h(s,I_0,\phi),I_0)ds,\qquad \Delta\theta_-(\phi)=\int_{-\infty}^0 \Omega(x^h(s,I_0,\phi),I_0)ds.$$

We are now ready to state the three main theorems of this paper. All three concern special orbits on the intersection surface $\Sigma_\varepsilon^\lambda(\bar\phi,\bar\theta_0)$. These orbits are homoclinic or heteroclinic connections between equilibria and periodic orbits in the resonance band, and arise in three different geometric situations. Orbits described in Theorems 1 and 3 exist for discrete values of the parameter λ, whereas orbits described in Theorem 2 exist on intervals of λ.

The first special situation that leads to the existence of homoclinic or heteroclinic orbits occurs when there exist two families of curves $O_{1,\varepsilon}(\lambda)$ and $O_{2,\varepsilon}(\lambda)$ on $\hat{\mathcal{M}}_\varepsilon$ in the resonance region for λ near $\bar\lambda$ and all small enough ε. The curve $O_{1,\varepsilon}(\lambda)$ can be either a stable periodic orbit for the restricted system (3.2) on $\hat{\mathcal{M}}_\varepsilon$ or a (restricted) unstable manifold of a saddle for this system. The curve $O_{2,\varepsilon}(\lambda)$ can be either an

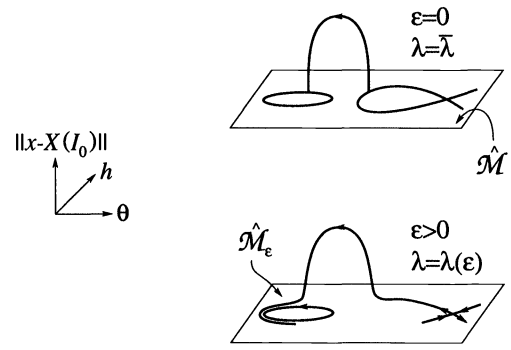


FIG. 4. *Here, the curve $O_{1,\varepsilon}((\lambda)$ is the restricted unstable manifold of the saddle $s_\varepsilon(\lambda)$ and the curve $O_{2,\varepsilon}(\lambda)$ is an unstable limit cycle on the annulus $\hat{\mathcal{M}}_\varepsilon$, as discussed in Theorem 1. In this case, a heteroclinic orbit connects the saddle $s_\varepsilon(\lambda(\varepsilon))$ to the periodic orbit $O_{2,\varepsilon}(\lambda(\varepsilon))$ for some $\lambda = \lambda(\varepsilon)$.*

unstable periodic orbit for the restricted system (3.2) on $\hat{\mathcal{M}}_\varepsilon$ or a (restricted) stable manifold of a saddle for this system.

To set up the geometry for the first theorem, shown in Fig. 4, let for $\lambda = \bar{\lambda}$ the line $\theta = \bar{\theta}_0 - \Delta\theta_-(\bar{\phi})$ intersect transversely the curve $O_{1,0}(\bar{\lambda})$, and the line $\theta = \bar{\theta}_0 + \Delta\theta_+(\bar{\phi})$ intersect transversely the curve $O_{2,0}(\bar{\lambda})$, and let both intersections occur at the same value of h. In this case a heteroclinic orbit on the limiting homoclinic surface $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$ connects the two intersection points. Assume further that, for $\lambda > \bar{\lambda}$, the h-coordinate of the intersection of the line $\theta = \bar{\theta}_0(I_0, \lambda) - \Delta\theta_-(\bar{\phi}(I_0, \lambda))$ and the curve $O_{1,0}(\lambda)$ is larger (smaller) than the h-coordinate of the intersection of the line $\theta = \bar{\theta}_0(I_0, \lambda) + \Delta\theta_+(\bar{\phi}(I_0, \lambda))$ and the curve $O_{2,0}(\lambda)$. Assume also that for $\lambda < \bar{\lambda}$, the h-coordinate of the intersection of the line $\theta = \bar{\theta}_0(I_0, \lambda) - \Delta\theta_-(\bar{\phi}(I_0, \lambda))$ and the curve $O_{1,0}(\lambda)$ is smaller (larger) than the h-coordinate of the intersection of the line $\theta = \bar{\theta}_0(I_0, \lambda) + \Delta\theta_+(\bar{\phi}(I_0, \lambda))$ and the curve $O_{2,0}(\lambda)$. See Fig. 5. In other words, the difference of these h-coordinates passes through zero transversely as λ passes through $\bar{\lambda}$. We will then show that for some $\lambda = \lambda(\varepsilon)$, there exists a heteroclinic connection between the orbits $O_{1,\varepsilon}(\lambda(\varepsilon))$ and $O_{2,\varepsilon}(\lambda(\varepsilon))$, for all small enough ε.

We now proceed to formalize this discussion. We begin by denoting the h-coordinates of the two intersections discussed in the previous paragraph by $h\left(\bar{\theta}_0(I_0, \lambda) - \Delta\theta_-(\bar{\phi}(I_0, \lambda))\right)$ and $h\left(\bar{\theta}_0(I_0, \lambda) + \Delta\theta_+(\bar{\phi}(I_0, \lambda))\right)$, respectively. Rather than to calculate the difference of these two h-coordinates, it is more convenient to calculate their squares, using equation (3.4). Thus,

$$\frac{1}{2}h^2\left(\bar{\theta}_0(I_0, \lambda) - \Delta\theta_-(\bar{\phi}(I_0, \lambda))\right) = \mathcal{H}\left(O_{1,0}(\lambda)\right) - V\left(\bar{\theta}_0(I_0, \lambda) - \Delta\theta_-(\bar{\phi}(I_0, \lambda)), \lambda\right)$$

and

$$\frac{1}{2}h^2\left(\bar{\theta}_0(I_0, \lambda) + \Delta\theta_+(\bar{\phi}(I_0, \lambda))\right) = \mathcal{H}\left(O_{2,0}(\lambda)\right) - V\left(\bar{\theta}_0(I_0, \lambda) + \Delta\theta_+(\bar{\phi}(I_0, \lambda)), \lambda\right),$$

where $\mathcal{H}\left(O_{1,0}(\lambda)\right)$ and $\mathcal{H}\left(O_{2,0}(\lambda)\right)$ are the respective values of the rescaled Hamiltonian (3.4) on the curves $O_{1,0}(\lambda)$ and $O_{2,0}(\lambda)$. Then, we have the following theorem.

THEOREM 1. *Let the curve $O_{1,\varepsilon}(\lambda)$ be either a stable periodic orbit for the restricted system (3.2) on $\hat{\mathcal{M}}_\varepsilon$ or a (restricted) unstable manifold of a saddle $s_{1,\varepsilon}(\lambda)$ for this system, and let the curve $O_{2,\varepsilon}(\lambda)$ be either an unstable periodic orbit for the*

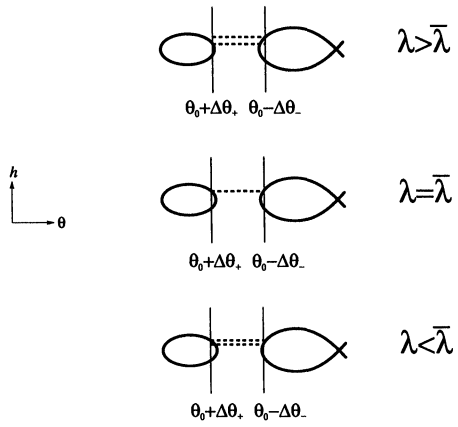


FIG. 5. *The transversality condition needed for Theorem* 1 *to be valid.*

restricted system (3.2) *on* $\hat{\mathcal{M}}_\varepsilon$ *or a (restricted) stable manifold of a saddle* $s_{2,\varepsilon}(\lambda)$ *for this system. Moreover, let for* $\lambda = \bar{\lambda}$, *the line* $\theta = \bar{\theta}_0 - \Delta\theta_-(\bar{\phi})$ *intersect transversely the curve* $O_{1,0}(\bar{\lambda})$ *and the line* $\theta = \bar{\theta}_0 + \Delta\theta_+(\bar{\phi})$ *intersect transversely the curve* $O_{2,0}(\bar{\lambda})$. *Finally, let*

$$\begin{aligned} &\mathcal{H}\left(O_{1,0}(\bar{\lambda})\right) - V\left(\bar{\theta}_0 - \Delta\theta_-(\bar{\phi}), \bar{\lambda}\right) \\ &\qquad - \left[\mathcal{H}\left(O_{2,0}(\bar{\lambda})\right) - V\left(\bar{\theta}_0 + \Delta\theta_+(\bar{\phi}), \bar{\lambda}\right)\right] = 0, \end{aligned} \tag{3.7}$$

and

$$\begin{aligned} &\frac{d}{d\lambda}\{\mathcal{H}\left(O_{1,0}(\lambda)\right) - V\left(\bar{\theta}_0(I_0,\lambda) - \Delta\theta_-(\bar{\phi}(I_0,\lambda)), \lambda\right) \\ &\qquad - \left[\mathcal{H}\left(O_{2,0}(\lambda)\right) - V\left(\bar{\theta}_0(I_0,\lambda) + \Delta\theta_+(\bar{\phi}(I_0,\lambda)), \lambda\right)\right]\} \neq 0. \end{aligned} \tag{3.8}$$

Then, for all small enough ε *and for some* $\lambda = \lambda(\varepsilon)$ *with* $\lambda(0) = \bar{\lambda}$, *there exists a heteroclinic orbit connecting either the periodic orbit* $O_{1,\varepsilon}(\lambda(\varepsilon))$ *or the saddle* $s_{1,\varepsilon}(\lambda)$ *to either the periodic orbit* $O_{2,\varepsilon}(\lambda(\varepsilon))$ *or the saddle* $s_{2,\varepsilon}(\lambda)$.

We remark that a saddle of the system (3.2) must always exist near a saddle of the outer system (3.3) by Proposition 5.3, and that a sufficient condition for the existence of limit cycles in the phase plane of the system (3.2) and a criterion for their stability is given in Proposition 5.5.

We next turn our attention to the situation that involves a heteroclinic connection between a saddle or a stable limit cycle on the annulus $\hat{\mathcal{M}}_\varepsilon$ and a sink or another stable limit cycle on $\hat{\mathcal{M}}_\varepsilon$, such as the one shown in Fig. 6. This situation is described in Theorem 2.

THEOREM 2. *Let the curve* $O_{1,\varepsilon}(\lambda)$ *be either a stable periodic orbit on the annulus* $\hat{\mathcal{M}}_\varepsilon$, *or the (restricted) unstable manifold of a saddle* $s_\varepsilon(\lambda)$ *on the annulus* $\hat{\mathcal{M}}_\varepsilon$ *for all* λ *near* $\lambda = \bar{\lambda}$, *and all small enough positive* ε. *At* $\lambda = \bar{\lambda}$, *let the curve* $O_{1,0}(\bar{\lambda})$ *and the line* $\theta = \bar{\theta}_0 - \Delta\theta_-(\bar{\phi})$ *intersect transversely at some height* $h = \bar{h}$. *Furthermore, let the point* $(h,\theta) = \left(\bar{h}, \bar{\theta}_0 + \Delta\theta_+(\bar{\phi})\right)$ *lie in a compact domain* $\mathcal{R}$ *that is all contained in the open region* $\mathcal{B}$, *the limit as* $\varepsilon \to 0$ *of the basin of attraction* $\mathcal{B}_\varepsilon$ *of either an equilibrium,*

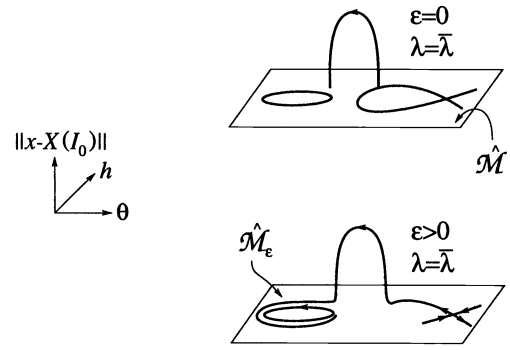


FIG. 6. *Here, a heteroclinic orbit connects the saddle* $s_\varepsilon(\lambda(\varepsilon))$ *to the periodic orbit* $O_\varepsilon((\lambda(\varepsilon))$, *which is a stable limit cycle on the annulus* $\hat{\mathcal{M}}_\varepsilon$, *as discussed in Theorem* 2.

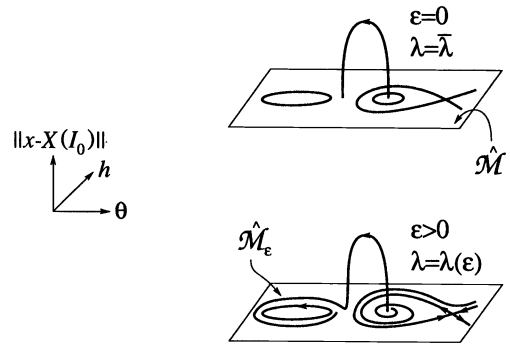


FIG. 7. *As discussed in Theorem* 3, *a heteroclinic orbit connects the spiral-saddle* $c_\varepsilon(\lambda(\varepsilon))$ *to the periodic orbit* $O_\varepsilon((\lambda(\varepsilon))$. *This periodic orbit is a stable limit cycle on the annulus* $\hat{\mathcal{M}}_\varepsilon$.

$c_\varepsilon(\lambda)$, *which is a sink for the restricted system* (3.2) *on the perturbed annulus* $\hat{\mathcal{M}}_\varepsilon$, *or a periodic orbit,* $O_{2,\varepsilon}(\lambda)$, *which is a stable limit cycle for the restricted system* (3.2) *on the perturbed annulus* $\hat{\mathcal{M}}_\varepsilon$. *Then, for all small enough positive* ε *and all* λ *close enough to* $\lambda = \bar{\lambda}$, *there exists a heteroclinic orbit connecting either the periodic orbit* $O_{1,\varepsilon}(\lambda)$ *or the saddle* $s_\varepsilon(\lambda)$ *to either the equilibrium* $c_\varepsilon(\lambda)$ *or the periodic orbit* $O_{2,\varepsilon}(\lambda)$. *Moreover, the intersection of the unstable manifolds* $W^{\mathrm{u}}(O_{1,\varepsilon}(\lambda))$ *or* $W^{\mathrm{u}}(s_\varepsilon(\lambda))$ *with the stable manifolds* $W^{\mathrm{s}}(c_\varepsilon(\lambda))$ *or* $W^{\mathrm{s}}(O_{2,\varepsilon}(\lambda))$ *is transverse along that heteroclinic orbit.*

Sufficient conditions for the existence of sources, sinks, and limit cycles in the phase plane of system (3.2) and criteria for their stability are given in Propositions 5.3 and 5.5, respectively. The existence of the limiting region $\mathcal{B}$ is a part of the assumption, and needs to be checked in each practical case separately. Also, by inverting the time, we can use this theorem to show the existence of a heteroclinic connection between a saddle or an unstable limit cycle on the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$ and a source or another unstable limit cycle on that annulus.

Finally, we discuss the situation, which involves a heteroclinic connection between two equilibria that are sinks for the restricted system (3.2), or a sink and a stable periodic orbit for that system; see Fig. 7. This situation is described in Theorem 3.

THEOREM 3. *Let* $c_0(\lambda)$ *be a center for the outer system* (3.3), *and let it be, at*

$\lambda = \bar{\lambda}$, *located at*

$$(h(c_0(\bar{\lambda})), \theta(c_0(\bar{\lambda}))) = (0, \bar{\theta}_0 - \Delta\theta_-(\bar{\phi})), \tag{3.9}$$

with

$$\frac{d}{d\lambda}\left[\theta(c_0(\lambda)) - \bar{\theta}_0(I_0, \lambda) + \Delta\theta_-(\bar{\phi}(I_0, \lambda))\right] \neq 0 \tag{3.10}$$

at $\lambda = \bar{\lambda}$. *Let the corresponding perturbed equilibrium* $c_\varepsilon(\lambda)$ *be a sink for the restricted system* (3.2) *for all small enough* ε *and all* λ *close enough to* $\lambda = \bar{\lambda}$. *Finally, let the point* $(h, \theta) = (0, \bar{\theta}_0 + \Delta\theta_+(\bar{\phi}))$ *lie in a compact domain* $\mathcal{R}$ *that is all contained in the open region* $\mathcal{B}$, *the limit as* $\varepsilon \to 0$ *of the basin of attraction* $\mathcal{B}_\varepsilon$ *of either an equilibrium,* $d_\varepsilon(\lambda)$, *which is a sink for the restricted system* (3.2) *on the perturbed annulus* $\hat{\mathcal{M}}_\varepsilon$, *or a periodic orbit,* $O_\varepsilon(\lambda)$, *which is a stable limit cycle for the restricted system* (3.2) *on the perturbed annulus* $\hat{\mathcal{M}}_\varepsilon$. *Then, for small* $\varepsilon > 0$, *there exists a function* $\lambda = \lambda(\varepsilon)$ *with* $\lambda(0) = \bar{\lambda}$, *such that there exists a heteroclinic orbit connecting the equilibrium* $c_\varepsilon(\lambda(\varepsilon))$ *to either the equilibrium* $d_\varepsilon(\lambda(\varepsilon))$ *or the periodic orbit* $O_\varepsilon(\lambda(\varepsilon))$.

This theorem is a slight generalization of the result of [30]. We remark that, again by inverting the time, we can use Theorem 3 to show the existence of a heteroclinic connection between two equilibria that are sources for the restricted system (3.2), or a source and an unstable limit cycle for (3.2).

The rest of the paper is devoted to the necessary background and the proofs of Theorems 1–3.

4. Homoclinic intersection surfaces. In this section we first discuss the persistence of the annulus $\mathcal{M}$ and its stable and unstable manifolds, $W^s(\mathcal{M})$ and $W^u(\mathcal{M})$, for nonzero ε. We then review those features of the Melnikov method [1]–[17] that are necessary for understanding the rest of this paper. In particular, we focus our attention on calculating when the stable and unstable manifolds, $W^s(\mathcal{M}_\varepsilon)$ and $W^u(\mathcal{M}_\varepsilon)$, of the perturbed annulus $\mathcal{M}_\varepsilon$ intersect transversely, and what the nature of those intersections is. Alternatively, we can ask ourselves which homoclinic orbits will survive under perturbation.

To show persistence under perturbation of the annulus $\mathcal{M}$ and its stable and unstable manifolds, we will have to perform local analysis around the annulus $\mathcal{M}$, and hence we now define the local stable and unstable manifold of $\mathcal{M}$. We pick a small positive δ and choose a neighborhood

$$U_\delta = \{(x, I, \theta) \mid I_1 < I < I_2, \|x - X(I)\| < \delta, 0 \leq \theta \leq 2\pi\}$$

of the annulus $\mathcal{M}$. We define the *local stable manifold*, $W^s_{\text{loc}}(\mathcal{M})$, of $\mathcal{M}$ to be the component of $W^s(\mathcal{M}) \cap U_\delta$ whose points do not leave the neighborhood U_δ in forward time. Thus, the local stable manifold $W^s_{\text{loc}}(\mathcal{M})$ consists precisely of those points in the neighborhood U_δ which asymptote toward the annulus $\mathcal{M}$ in positive time without ever leaving U_δ. We define $W^u_{\text{loc}}(\mathcal{M})$, the *local unstable manifold* of the annulus $\mathcal{M}$, in an analogous fashion. If κ is any number smaller than $\inf\{\kappa(I) \mid I_1 < I < I_2\}$, where $\kappa(I)$ is the smallest positive real part of the eigenvalues of the matrix $JD^2_x H(X(I), I)$, then trajectories on the local stable manifold $W^s_{\text{loc}}(\mathcal{M})$ approach the annulus $\mathcal{M}$ in forward time exponentially at least at the rate $e^{-\kappa t}$. A similar statement is true for trajectories on the local unstable manifold $W^u_{\text{loc}}(\mathcal{M})$ in backward time.

We are now ready to state the precise result that describes how the annulus $\mathcal{M}$ and its local stable and unstable manifolds persist under perturbation.

PROPOSITION 4.1. *For all small enough positive ε, there exist a two-dimensional, locally invariant annular surface, $\mathcal{M}_\varepsilon$, and $(n+2)$-dimensional, locally invariant manifolds, $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ and $W^u_{\rm loc}(\mathcal{M}_\varepsilon)$, inside the neighborhood U_δ, possessing the following properties:*

1. *The annulus $\mathcal{M}_\varepsilon$ and the manifolds $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ and $W^u_{\rm loc}(\mathcal{M}_\varepsilon)$ vary smoothly with ε and other parameters in the problem.*

2. *For $\varepsilon = 0$, the annulus $\mathcal{M}_\varepsilon$ and the manifolds $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ and $W^u_{\rm loc}(\mathcal{M}_\varepsilon)$ coincide with the annulus $\mathcal{M}$ and the manifolds $W^s_{\rm loc}(\mathcal{M})$ and $W^u_{\rm loc}(\mathcal{M})$, respectively. For nonzero ε, the annulus $\mathcal{M}_\varepsilon$ and the manifolds $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ and $W^u_{\rm loc}(\mathcal{M}_\varepsilon)$ can be written as smooth graphs over their unperturbed counterparts. In particular, the annulus $\mathcal{M}_\varepsilon$ is given by the equation $x = X_\varepsilon(I,\theta,\lambda,\varepsilon)$ for some smooth function $X_\varepsilon(I,\theta,\lambda,\varepsilon)$ with $X_0(I,\theta,\lambda,0) = X(I)$.*

3. *The manifolds $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ and $W^u_{\rm loc}(\mathcal{M}_\varepsilon)$ intersect along the annulus $\mathcal{M}_\varepsilon$.*

4. *Let κ be any number smaller than $\inf\{\kappa(I) \mid I_1 < I < I_2\}$, where $\kappa(I)$ is the smallest positive real part of the eigenvalues of the matrix $JD^2_x H(X(I),I)$. Then any trajectory that starts at $t = 0$ inside the manifold $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ will approach the annulus $\mathcal{M}_\varepsilon$ in forward time at an exponential rate at least as fast as $e^{-\kappa t}$, as long as it stays in the manifold $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$. Any trajectory that starts at $t = 0$ inside the manifold $W^u_{\rm loc}(\mathcal{M}_\varepsilon)$ will approach the annulus $\mathcal{M}_\varepsilon$ in backward time at an exponential rate at least as fast as $e^{\kappa t}$, as long as it stays in the manifold $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$.*

Proof. The proof of this proposition follows from [44]–[46]. Details are similar to those in [17, p. 354]. □

We call $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ and $W^u_{\rm loc}(\mathcal{M}_\varepsilon)$ the *local stable and unstable manifolds* of the perturbed annulus $\mathcal{M}_\varepsilon$, respectively. The meaning of the statement that the annulus $\mathcal{M}_\varepsilon$ and its local stable and unstable manifolds $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ and $W^u_{\rm loc}(\mathcal{M}_\varepsilon)$ are locally invariant is that they are spanned by orbits, but points can enter or leave them through their boundaries.

We define the *stable manifold*, $W^s(\mathcal{M}_\varepsilon)$, of the perturbed annulus $\mathcal{M}_\varepsilon$ as the manifold obtained by evolving points on the local stable manifold $W^s_{\rm loc}(\mathcal{M}_\varepsilon)$ in backward time. We note that trajectories can leave (but not enter) the stable manifold $W^s(\mathcal{M}_\varepsilon)$ through its boundary, which is enough to make the manifold $W^s(\mathcal{M}_\varepsilon)$ only locally invariant. This fact is in contrast with the usual properties of the stable manifold of an invariant manifold, which is itself invariant, and comes about because the perturbed annulus $\mathcal{M}_\varepsilon$ itself is only locally invariant. An analogous definition and comments hold for the *unstable manifold*, $W^u(\mathcal{M}_\varepsilon)$, of the perturbed annulus $\mathcal{M}_\varepsilon$.

Gronwall-type estimates yield Proposition 4.2.

PROPOSITION 4.2. *Two trajectories, one on the unperturbed stable manifold $W^s(\mathcal{M})$ and the other on the perturbed stable manifold $W^s(\mathcal{M}_\varepsilon)$, which start a distance $\mathcal{O}(\varepsilon)$ apart at $t = 0$, will stay $\mathcal{O}(\varepsilon)$ close for all finite times. The same statement also holds for pairs of trajectories on the unperturbed unstable manifold $W^u(\mathcal{M})$ and the perturbed unstable manifold $W^u(\mathcal{M}_\varepsilon)$.*

We now turn our attention to the question of which unperturbed homoclinic orbit will survive under perturbation. As mentioned in §3, the answer to this question is determined by the transverse zeros of the Melnikov vector (2.6). In particular, the equation $\mathbf{M}(I,\phi,\theta_0,\lambda) = 0$ presents n constraints on $(n+2)$ variables, I, ϕ, θ_0, and λ. Hence, if we fix the parameter λ, we expect this equation to provide a one-parameter family of surviving homoclinic orbits or, in other words, a two-dimensional homoclinic intersection surface. This discussion is made precise in Proposition 4.3.

PROPOSITION 4.3. *Let for $I = \bar{I}$, $\phi = \bar{\phi}$, $\theta_0 = \bar{\theta}_0$, and $\lambda = \bar{\lambda}$ the following two*

statements be true:

1. $\mathbf{M}(\bar{I}, \bar{\phi}, \bar{\theta}_0, \bar{\lambda}) = 0$.
2. *The matrix* $D_{(\phi,\theta_0)}\mathbf{M}(\bar{I}, \bar{\phi}, \bar{\theta}_0, \bar{\lambda})$ *is nonsingular.*

Then for ε *sufficiently small, all* I *sufficiently close to* $\bar{I}$*, and all* λ *sufficiently close to* $\bar{\lambda}$*, there exists a two-dimensional intersection surface* $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ *of the manifolds* $W^s(\mathcal{M}_\varepsilon)$ *and* $W^u(\mathcal{M}_\varepsilon)$ *in the* $x-I-\theta$ *phase space. The intersection surface* $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ *varies smoothly with* ε *and* λ*. In the limit as* $\varepsilon \to 0$*, it tends to the surface* $\Sigma^\lambda(\bar{\phi}, \bar{\theta}_0)$ *that consists of the unperturbed homoclinic orbits parametrized by* t *and* I*, with* $\phi = \bar{\phi}(I, \lambda)$*, and* $\theta_0 = \bar{\theta}_0(I, \lambda)$*, for* I *close enough to* $\bar{I}$ *at the given* λ*. The manifolds* $W^s(\mathcal{M}_\varepsilon)$ *and* $W^u(\mathcal{M}_\varepsilon)$ *intersect transversely at every point of the intersection surface* $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$.

By Assumption 5, the hypotheses of this proposition are satisfied for $I = I_0$ and some $\phi = \bar{\phi}$, $\theta_0 = \bar{\theta}_0$, and $\lambda = \bar{\lambda}$.

Proof of Proposition 4.3. Let $\mathbf{M}(\bar{I}, \bar{\phi}, \bar{\theta}_0, \bar{\lambda}) = 0$, and let the matrix of partial derivatives $D_{(\phi,\theta)}\mathbf{M}(\bar{I}, \bar{\phi}, \bar{\theta}_0, \bar{\lambda})$ be nonsingular. Then, by the implicit function theorem, there exist functions $\phi = \bar{\phi}(I, \lambda)$, $\theta_0 = \bar{\theta}_0(I, \lambda)$ with $\bar{\phi}(\bar{I}, \bar{\lambda}) = \bar{\phi}$ and $\bar{\theta}_0(\bar{I}, \bar{\lambda}) = \bar{\theta}_0$, such that $\mathbf{M}(I, \bar{\phi}(I, \lambda), \bar{\theta}_0(I, \lambda), \lambda) = 0$ and that $D_{(\phi,\theta)}\mathbf{M}(I, \bar{\phi}(I, \lambda), \bar{\theta}_0(I, \lambda), \lambda)$ has maximal rank in some small neighborhood of the point $(\bar{I}, \bar{\lambda})$ in the $I - \lambda$ space. The proof now follows from Theorems 4.1.9 and 4.1.10 in [17]. □

We now discuss the nature of the orbits contained in the intersections of the manifolds $W^s(\mathcal{M}_\varepsilon)$ and $W^u(\mathcal{M}_\varepsilon)$. We remark that even though a trajectory on an intersection orbit is $\mathcal{O}(\varepsilon)$ close to the trajectory given by $\big(x^h(t, I, \bar{\phi}(I, \lambda)), I, \bar{\theta}^h(t, I, \bar{\phi}(I, \lambda)) + \bar{\theta}_0(I, \lambda)\big)$ for all finite $t \in \mathbb{R}$, the two orbits traced by these two trajectories are not necessarily uniformly close. Namely, by Proposition 4.2, the two trajectories are only guaranteed to be uniformly close on compact intervals of t. After that, they may move away from each other, since the usual Gronwall-type estimate (see, for instance, [33]) only bounds their distance as $\varepsilon C_1 e^{C_2|t|}$, where C_1 and C_2 are some appropriate positive constants.

We also note that an intersection orbit is contained in the stable manifold $W^s(\mathcal{M}_\varepsilon)$ forever in backward time and in the unstable manifold $W^u(\mathcal{M}_\varepsilon)$ forever in forward time by the very construction of those manifolds. However, in forward time, an intersection orbit may leave the stable manifold $W^s(\mathcal{M}_\varepsilon)$ through its boundary because this manifold is only locally invariant. It can also leave the unstable manifold $W^u(\mathcal{M}_\varepsilon)$ through its boundary in backward time for the same reason. Therefore, we expect most intersection orbits not to asymptote to any invariant object in the annulus $\mathcal{M}_\varepsilon$ in either forward or backward time. However, all the orbits that do asymptote to invariant objects in the perturbed annulus $\mathcal{M}_\varepsilon$ in both forward and backward time must be contained in the intersection of its stable and unstable manifolds, $W^s(\mathcal{M}_\varepsilon)$ and $W^u(\mathcal{M}_\varepsilon)$.

If $\bar{I} = I_0$, and if we set $I = I_0 + \sqrt{\varepsilon}h$ and let $\varepsilon \to 0$, it should be clear that the homoclinic intersection surface $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ tends, in the inner limit, to the limiting homoclinic intersection surface $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$, described in §3.

5. The resonance band. In this section we present a way to analyze the dynamics in the resonance band that emerges from the breakup of the circle of equilibria at $I = I_0$ for nonzero ε. We have seen in Proposition 4.1 that the x coordinates of points on the perturbed annulus $\mathcal{M}_\varepsilon$ are given by the equality $x = X_\varepsilon(I, \theta, \lambda, \varepsilon)$. Therefore, the following two equations completely describe the dynamics on the annulus $\mathcal{M}_\varepsilon$:

$$\dot{I} = \varepsilon g^I(X_\varepsilon(I, \theta, \lambda, \varepsilon), I, \theta, \lambda)\,, \tag{5.1a}$$

$$\dot{\theta} = \Omega(X_\varepsilon(I,\theta,\lambda,\varepsilon),I) + \varepsilon g^\theta(X_\varepsilon(I,\theta,\lambda,\varepsilon),I,\theta,\lambda). \tag{5.1b}$$

Since we are interested in the dynamics in the resonance band near $I = I_0$, we also substitute $I = I_0 + \sqrt{\varepsilon}h$ into equations (5.1). We remark that the scaling factor in front of h is $\sqrt{\varepsilon}$ because we have assumed in Assumption 3 the generic situation in which $\Omega(X(I_0),I_0) = 0$ and $\frac{d\Omega}{dI}(X(I_0),I_0) \neq 0$. Otherwise, this factor may be different; see [32]–[34].

We can only extract useful information about equations (5.1) near the resonance at $I = I_0$ if we can explicitly calculate, or at least approximate, the function $X_\varepsilon(I_0 + \sqrt{\varepsilon}h,\theta,\lambda,\varepsilon)$. This is indeed the case, as we now show. First, as a consequence of Proposition 4.1, the function $X_\varepsilon(I,\theta,\lambda,\varepsilon)$ is smooth, so that it can be Taylor expanded about $\varepsilon = 0$ as follows:

$$\begin{aligned} X_\varepsilon(I,\theta,\lambda,\varepsilon) &= X(I) + \varepsilon X_1(I,\theta,\lambda) + \varepsilon^2 X_2(I,\theta,\lambda) \\ &\quad + \cdots + \varepsilon^{m-1} X_{m-1}(I,\theta,\lambda) + \mathcal{O}(\varepsilon^m), \end{aligned} \tag{5.2}$$

where m is at most equal to the number of continuous derivatives of the vector field (2.1). Second, we can in fact calculate all the terms in the Taylor expansion in powers of $\sqrt{\varepsilon}$ of the function $X_\varepsilon(I_0 + \sqrt{\varepsilon}h,\theta,\lambda,\varepsilon)$, as is implied by the following result.

PROPOSITION 5.1. *At the resonant value $I = I_0$, the terms $X_i(I_0,\theta,\lambda)$ in the Taylor expansion* (5.2) *and their partial derivatives $D_I^j X_i(I_0,\theta,\lambda)$ can be calculated recursively in terms of differentiations and algebraic operations alone. In particular,*

$$X_1(I_0,\theta,\lambda) = \left[JD_x^2 H(X(I_0),I_0)\right]^{-1} \left(g^I(X(I_0),I_0,\theta,\lambda)\frac{dX(I_0)}{dI} - g^x(X(I_0),I_0,\theta,\lambda)\right). \tag{5.3}$$

Proof. Let $X_\varepsilon = X_\varepsilon(I,\theta,\lambda,\varepsilon)$ and proceed as in [37] and [30]. By equation (2.1a), we have

$$\dot{X}_\varepsilon = JD_x H(X_\varepsilon,I) + \varepsilon g^x(X_\varepsilon,I,\theta,\lambda).$$

On the other hand, we obtain by the chain rule and equations (2.1b) and (2.1c) the equation

$$\begin{aligned} \dot{X}_\varepsilon &= D_I X_\varepsilon \dot{I} + D_\theta X_\varepsilon \dot{\theta} \\ &= D_I X_\varepsilon \varepsilon g^I(X_\varepsilon,I,\theta,\lambda) + D_\theta X_\varepsilon \left(\Omega(X_\varepsilon,I) + \varepsilon g^\theta(X_\varepsilon,I,\theta,\lambda)\right). \end{aligned}$$

We equate the two expressions for $\dot{X}_\varepsilon$, Taylor expand using formula (5.2), and examine the $\mathcal{O}(\varepsilon^i)$ term for $i = 1,\ldots,m-1$.

When $i = 1$, we obtain the equation

$$\begin{aligned} &JD_x^2 H(X(I),I)X_1(I,\theta,\lambda) + g^x(X(I),I,\theta,\lambda) \\ &\qquad = g^I(X(I),I,\theta,\lambda)\frac{dX(I)}{dI} + \Omega(X(I),I)D_\theta X_1(I,\theta,\lambda). \end{aligned}$$

Formula (5.3) now follows upon setting $I = I_0$ because $\Omega(X(I_0),I_0) = 0$ by Assumption 3.

Similarly, for $i = 2, \ldots, m-1$, we obtain the equation

$$JD_x^2 H(X(I), I) X_i(I, \theta, \lambda)$$

$$= \Phi_i \Bigg(X(I), \frac{dX(I)}{dI}, X_1(I, \theta, \lambda), D_I X_1(I, \theta, \lambda), D_\theta X_1(I, \theta, \lambda), \ldots,$$

$$X_{i-1}(I, \theta, \lambda), D_I X_{i-1}(I, \theta, \lambda), D_\theta X_{i-1}(I, \theta, \lambda), I, \theta, \lambda \Bigg)$$

$$+\Omega(X(I), I) D_\theta X_i(I, \theta, \lambda), \tag{5.4}$$

for some smooth function Φ_i. Upon setting $I = I_0$, the expression for $X_i(I_0, \theta, \lambda)$ readily follows as above.

The expressions for the derivatives $D_\theta X_i(I_0, \theta, \lambda)$ can be computed by simply differentiating the formulas for $X_i(I_0, \theta, \lambda)$ with respect to θ, and so the statement about the derivatives $D_I^j X_i(I_0, \theta, \lambda)$ follows after differentiating formula (5.4), and setting $I = I_0$. □

In practice, the first-order corrections in the expansion about $\sqrt{\varepsilon} = 0$ of equations (5.1) with $I = I_0 + \sqrt{\varepsilon} h$ should suffice. Upon rescaling the time using $\tau = \sqrt{\varepsilon} t$ and setting $' = \frac{d}{d\tau}$, these equations read

$$h' = g^I(X(I_0), I_0, \theta, \lambda) + \sqrt{\varepsilon} F_h(h, \theta, \lambda) + \mathcal{O}(\varepsilon), \tag{5.5a}$$

$$\theta' = \frac{d\Omega}{dI}(X(I_0), I_0)\, h + \sqrt{\varepsilon} F_\theta(h, \theta, \lambda) + \mathcal{O}(\varepsilon), \tag{5.5b}$$

with

$$F_h(h, \theta, \lambda) = \frac{d}{dI} \left[g^I(X(I_0), I_0, \theta, \lambda) \right] h \tag{5.6}$$

and

$$F_\theta(h, \theta, \lambda) = \frac{1}{2} \frac{d^2 \Omega(X(I_0), I_0)}{dI^2} h^2 + D_x \Omega(X(I_0), I_0) X_1(I_0, \theta, \lambda)$$

$$+ g^\theta(X(I_0), I_0, \theta, \lambda), \tag{5.7}$$

where $X_1(I_0, \theta, \lambda)$ is given by formula (5.3).

As stated in §3, in the limit as $\varepsilon \to 0$, we obtain the outer system (3.3)

$$h' = g^I(X(I_0), I_0, \theta, \lambda), \qquad \theta' = \frac{d\Omega}{dI}(X(I_0), I_0)\, h.$$

Since the outer system has the special form of a one-degree-of-freedom Newtonian system with the Hamiltonian function (3.4),

$$\mathcal{H}(h, \theta, \lambda) = \frac{1}{2} \frac{d\Omega}{dI}(X(I_0), I_0)\, h^2 + V(\theta, \lambda)$$

$$= \frac{1}{2} \frac{d\Omega}{dI}(X(I_0), I_0)\, h^2 - \int_0^\theta g^I(X(I_0), I_0, s, \lambda) ds,$$

it is easy to analyze and possesses certain simple general properties. (See [47].) These properties are stated in Proposition 5.2.

PROPOSITION 5.2. *Orbits of the outer system* (3.3) *are symmetric about the* θ*-axis, and its equilibria can only lie on the* θ*-axis. Also, since the function* $g^I(x, I, \theta, \lambda)$ *is periodic in* θ*, the system* (3.3) *can only have an even number of equilibria for* θ *in* $[0, 2\pi)$*, provided that all the zeros of the expression* $g^I(X(I_0), I_0, \theta, \lambda)$ *are simple. Moreover, the derivative* $D_\theta g^I(X(I_0), I_0, \theta, \lambda)$ *must have opposite signs at two consecutive zeros* θ_1 *and* θ_2*. Therefore, one of any two neighboring equilibria of* (3.3) *must be a center and the other, a saddle.*

We notice that the potential part, $V(\theta, \lambda)$, of the Hamiltonian $\mathcal{H}(h, \theta, \lambda)$ can be written in the form

$$\begin{aligned} V(\theta, \lambda) &= \hat{V}(\theta, \lambda) - \bar{V}(\lambda)\theta \\ &= -\int_0^\theta \left[g^I\left(X(I_0), I_0, s, \lambda\right) - \bar{V}(\lambda)\right] ds - \bar{V}(\lambda)\theta, \end{aligned} \tag{5.8}$$

where

$$\bar{V}(\lambda) = \frac{1}{2\pi} \int_0^{2\pi} g^I(X(I_0), I_0, s, \lambda) ds\,.$$

The expression $\hat{V}(\theta, \lambda)$ in formula (5.8) is periodic in θ, and the expression $\bar{V}(\lambda)\theta$ is linear in θ. Formula (3.4) then shows that, say, if $\bar{V}(\lambda)$ and $\frac{d\Omega}{dI}(X(I_0), I_0)$ have the same sign, then every orbit on the $h-\theta$ phase cylinder must be bounded from the left. In particular, the left-hand halves of the stable and unstable manifolds of a saddle must either coincide along a homoclinic orbit or form two heteroclinic connections to another saddle. In both cases, the homoclinic orbit, or heteroclinic connections, encircle a center. A similar statement is true if $\bar{V}(\lambda)$ and $\frac{d\Omega}{dI}(X(I_0), I_0)$ have opposite signs.

We also notice that the decomposition (5.8) of the potential $V(\theta, \lambda)$ implies that the Hamiltonian $\mathcal{H}(h, \theta, \lambda)$ is not a single-valued function on the $h - \theta$ cylinder.

By Assumption 4, the function $g^I(X(I_0), I_0, \theta, \lambda)$ has only a finite number of simple zeros; that is, the partial derivative $D_\theta g^I(X(I_0), I_0, \theta, \lambda)$ is nonzero at each zero of $g^I(X(I_0), I_0, \theta, \lambda)$. The implicit function theorem immediately implies Proposition 5.3.

PROPOSITION 5.3. *Every equilibrium of the system* (3.3) *persists in the system* (3.2) *a distance* $\mathcal{O}(\sqrt{\varepsilon})$ *away (in the* $x - h - \theta$ *coordinates), and also persists for neighboring values of the parameter* λ*. If the unperturbed equilibrium is a saddle, so is the perturbed one. A sufficient condition for an unperturbed center to perturb into a source or a sink is that the expression* $D_h F_h(h, \theta, \lambda) + D_\theta F_\theta(h, \theta, \lambda)$ *calculated at that center be positive or negative, respectively.*

The last sentence in this proposition is true because the real parts of both eigenvalues at the perturbed counterpart of a center are equal to

$$\frac{\sqrt{\varepsilon}}{2} \left[D_h F_h(h, \theta, \lambda) + D_\theta F_\theta(h, \theta, \lambda)\right] + \mathcal{O}(\varepsilon).$$

Therefore, the center perturbs into a source on the cylinder $\mathcal{M}_\varepsilon$ if this expression is positive and into a sink if this expression is negative. (Recall that the expressions $F_h(h, \theta, \lambda)$ and $F_\theta(h, \theta, \lambda)$ are given by formulas (5.6) and (5.7), respectively.)

The usual stable manifold theorem and Gronwall-type estimates imply Proposition 5.4.

PROPOSITION 5.4. *Let s_0 be a saddle for the outer system* (3.3), *and let s_ε be its perturbed counterpart for small positive ε. Then, a trajectory on the unstable manifold of the restricted system* (3.2) *of the perturbed saddle s_ε is $\mathcal{O}(\sqrt{\varepsilon})$ close (in the $x-h-\theta$ coordinates) to a trajectory on the unstable manifold of the unperturbed saddle s_0 on the τ-interval $(-\infty, T]$ for all finite T. A similar statement holds for the stable manifolds of the saddles s_ε and s_0 on τ-intervals $[T, \infty)$.*

Periodic orbits of the outer system (3.3) may also survive in the system (3.2).

PROPOSITION 5.5. *A periodic orbit of the outer system* (3.3) *will survive in the system* (3.2) *if the subharmonic Melnikov function*

$$\begin{aligned} M(\mathcal{H}) &= \oint F_h(h,\theta,\lambda)d\theta - F_\theta(h,\theta,\lambda)dh \\ &= \oint \left[\frac{d\Omega}{dI}(X(I_0),I_0)h(\tau)F_h(h(\tau),\theta(\tau),\lambda)\right. \\ &\qquad \left. -g^I(X(I_0),I_0,\theta(\tau),\lambda)F_\theta(h(\tau),\theta(\tau),\lambda)\right]d\tau, \end{aligned}$$

calculated around that periodic orbit, has a transverse zero as a function of $\mathcal{H}$, the orbit's energy in the system (3.3). *This orbit then also persists for neighboring λ. Moreover, if $dM(\mathcal{H})/d\mathcal{H}$ is positive on the persisting orbit, then this orbit is unstable on the perturbed annulus $\mathcal{M}_\varepsilon$; and if $dM(\mathcal{H})/d\mathcal{H}$ is negative on the persisting orbit, then this orbit is stable.*

For a special case, a proof of this proposition is given in [47, p. 92]. The more general case described here is proven in the same way.

6. Geometric singular perturbation theory. To study the orbit structure on the stable and unstable manifolds $W^s(\mathcal{M}_\varepsilon)$ and $W^u(\mathcal{M}_\varepsilon)$ near the perturbed annulus $\mathcal{M}_\varepsilon$ and near the resonance at $I = I_0$, we use the rescaling $I = I_0 + \sqrt{\varepsilon}\,h$ on the full system (2.1) to obtain the system

$$\dot{x} = JD_xH(x, I_0 + \sqrt{\varepsilon}\,h) + \varepsilon g^x(x, I_0 + \sqrt{\varepsilon}\,h, \theta, \lambda)\,, \tag{6.1a}$$

$$\dot{h} = \sqrt{\varepsilon}\,g^I(x, I_0 + \sqrt{\varepsilon}h, \theta, \lambda)\,, \tag{6.1b}$$

$$\dot{\theta} = \Omega\,(x, I_0 + \sqrt{\varepsilon}\,h) + \varepsilon g^\theta(x, I_0 + \sqrt{\varepsilon}\,h, \theta, \lambda)\,. \tag{6.1c}$$

Setting $\varepsilon = 0$ in equations (6.1), we obtain the inner system (3.5)

$$\dot{x} = JD_xH(x, I_0), \quad \dot{h} = 0, \quad \dot{\theta} = \Omega\,(x, I_0).$$

The properties of its phase space are described in §3. On the other hand, if we restrict the dynamics of equations (6.1) to the annulus $\mathcal{M}_\varepsilon$ and rescale the time into $\tau = \sqrt{\varepsilon}t$, we obtain in the limit as $\varepsilon \to 0$ the outer system (3.3)

$$h' = g(X(I_0), I_0, \theta, \lambda), \qquad \theta' = \frac{d\Omega(X(I_0), I_0)}{dI}\,h,$$

with $' = \frac{d}{d\tau}$. We studied this system in the previous section.

As we have already mentioned in §3, the inner system (3.5) describes the dynamics away from the limiting annulus $\hat{\mathcal{M}}$ at $x = X(I_0)$, and the outer system (3.3) describes the rescaled slow dynamics on the annulus $\hat{\mathcal{M}}$. The work of Fenichel [37], [44]–[46] provides a means to connect the dynamics of systems (3.5) and (3.3) to achieve a description of the orbits in the local stable and unstable manifolds $W^{\mathrm{s}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ and $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ of the annulus $\hat{\mathcal{M}}_\varepsilon$, the perturbed counterpart of the annulus $\hat{\mathcal{M}}$. First, Proposition 4.1 applies to the system (6.1), and provides for smooth dependence of the annulus $\hat{\mathcal{M}}_\varepsilon$ and its stable and unstable manifolds $W^{\mathrm{s}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ and $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ on ε up to and including $\varepsilon = 0$. Second, Theorem 9.1 in [37] applies to systems of the same type as (6.1) and guarantees that the local manifolds $W^{\mathrm{s}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ and $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ are foliated by *stable and unstable fibers.* The statement of this theorem, which is tailored to the needs of the present paper, is given in the next proposition. (See also [40].) The estimates in this proposition are stated in the $x - h - \theta$ coordinates. The proposition is given for stable fibers; the proposition for unstable fibers is the same except for obvious changes.

PROPOSITION 6.1. *For all small enough ε, the local stable manifold $W^{\mathrm{s}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ of the invariant annulus $\hat{\mathcal{M}}_\varepsilon$ is foliated by a family of disjoint n-dimensional manifolds called* stable fibers. *These stable fibers have the following additional properties:*

1. *They form a* locally positively invariant family; *that is, the image (under the forward-time flow) of any stable fiber is contained in a stable fiber as long as this image is contained in the local stable manifold $W^{\mathrm{s}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$.*

2. *Each stable fiber pierces the annulus $\hat{\mathcal{M}}_\varepsilon$ transversely inside the manifold $W^{\mathrm{s}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ in precisely one point, called its* base point.

3. *As the base points on the annulus $\hat{\mathcal{M}}_\varepsilon$ move under the dynamics of the vector field* (6.1), *the stable fibers move along with their base points and contract exponentially toward their base points in forward time as long as the base points stay in $\hat{\mathcal{M}}_\varepsilon$. The rate of this exponential contraction is at least $e^{-\kappa t}$, where κ is any number smaller than the smallest positive real part $\kappa(I_0)$ of all the eigenvalues of the equation* (2.2a) *at $I = I_0$ (or, equivalently, the inner equation* (3.5a)) *linearized around $x = X(I_0)$.*

4. *The family of stable fibers varies smoothly with $\sqrt{\varepsilon}$, λ, and any other parameters in the problem.*

5. *For $\varepsilon = 0$, that is, for system* (3.5), *the stable fibers are precisely the local stable manifolds of the equilibria on the $h - \theta$ annulus $\hat{\mathcal{M}}$.*

Theorems similar to Proposition 6.1 are also stated in [30], [38], [48].

Fenichel's fibers make it possible to construct local stable and unstable manifolds of periodic orbits and equilibria on the annulus $\hat{\mathcal{M}}_\varepsilon$.

PROPOSITION 6.2. *The local stable and unstable manifolds of objects on the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$ are characterized by the following properties:*

1. *For every stable periodic orbit on the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$, its local unstable manifold is the union of all the unstable fibers whose base points lie on that orbit. The local part (contained in the neighborhood U_δ) of its stable manifold is the union of all the stable fibers whose base points are contained in the forward-time basin of attraction, $\mathcal{B}_\varepsilon$, of this periodic orbit on the annulus $\hat{\mathcal{M}}_\varepsilon$.*

2. *For every unstable periodic orbit on the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$, its local stable manifold is the union of all the stable fibers whose base points lie on that orbit. The local part (contained in the neighborhood U_δ) of its unstable manifold is the union of all the unstable fibers whose base points are contained in the backward-time basin of attraction, $\mathcal{B}_\varepsilon$, of this periodic orbit on the annulus $\hat{\mathcal{M}}_\varepsilon$.*

3. *For every equilibrium in the resonance that is a sink for the system* (6.1)

restricted to the annulus $\hat{\mathcal{M}}_\varepsilon$, that is, equations (3.2), *its local unstable manifold is precisely the unstable fiber having this sink as its base point. The local part (contained in the neighborhood U_δ) of its stable manifold is the union of all the stable fibers whose base points are contained in the forward-time basin of attraction, $\mathcal{B}_\varepsilon$, of this sink on the annulus $\hat{\mathcal{M}}_\varepsilon$.*

4. *For every equilibrium in the resonance that is a source for the equations* (3.2), *its local stable manifold is precisely the stable fiber having this source as its base point. The local part (contained in the neighborhood U_δ) of its unstable manifold is the union of all the unstable fibers whose base points are contained in the backward-time basin of attraction, $\mathcal{B}_\varepsilon$, of this source on the annulus $\hat{\mathcal{M}}_\varepsilon$.*

5. *For every equilibrium that is a saddle for the restricted system* (3.2), *the local parts (that lie in the neighborhood U_δ) of its stable and unstable manifolds are the unions of the stable and unstable fibers with base points lying on the restricted stable and unstable manifolds of this saddle on the annulus $\hat{\mathcal{M}}_\varepsilon$.*

This proposition is similar to Theorems 12.1, 12.2, 13.1, and 13.2 in [37].

We obtain the full global stable and unstable manifolds of orbits and equilibria that lie on the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$ by evolving their local counterparts in forward and backward time, respectively.

Using both systems of equations, (3.3) and (3.5), we can obtain the limiting structure of selected objects on the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$ and their stable and unstable manifolds. In particular, let O_ε be an orbit on $\hat{\mathcal{M}}_\varepsilon$ that limits, as $\varepsilon \to 0$, onto a curve O_0. This curve is a level curve of the rescaled Hamiltonian $\mathcal{H}(h,\theta)$ at some value $\mathcal{H}(h,\theta) = \mathcal{H}_0$. The curve O_0 is an orbit for the outer equations (3.3) and a curve of equilibria for the inner equations (3.5).

If O_ε is a periodic orbit, it is clear what is meant by its local stable and unstable manifolds. However, let O_ε be a piece of the unstable manifold of a saddle for the restricted system (3.2) on the annulus $\hat{\mathcal{M}}_\varepsilon$, parametrized by a trajectory on a τ-interval $(-\infty, T]$ for some large positive T. In this case, we define the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_\varepsilon)$ to be the union of all the unstable fibers whose base points lie on the curve O_ε. By Proposition 5.4, there exists a piece of the stable manifold of a saddle for the limiting outer system 3.3 on $\hat{\mathcal{M}}$ that is also parametrized by a trajectory on the τ-interval $(-\infty, T]$ and is $\mathcal{O}(\sqrt{\varepsilon})$ close to O_ε. This piece is the limiting curve O_0. An analogous definition can be given for the stable manifold $W^{\mathrm{s}}_{\mathrm{loc}}(O_\varepsilon)$ if the orbit O_ε is a piece of the stable manifold of a saddle for the restricted system (3.2).

The local stable and unstable manifolds, $W^{\mathrm{s}}_{\mathrm{loc}}(O_0)$ and $W^{\mathrm{u}}_{\mathrm{loc}}(O_0)$, of the limiting curve O_0 are the unions of the n-dimensional stable and unstable manifolds of the equilibria (under the dynamics of the inner equations (3.5)) that make up the curve O_0, respectively. The global stable and unstable manifolds of all the above-mentioned objects can now be defined in the usual way by evolving trajectories on the local stable and unstable manifolds in backward and forward time, respectively.

Propositions 6.1 and 6.2 now imply (in the $x - h - \theta$ coordinates) the following proposition, whose contents are illustrated in Figs. 8 and 9.

PROPOSITION 6.3. *In the limit when $\varepsilon \to 0$, the following limiting structures can be constructed with the aid of stable and unstable fibers:*

1. *If O_ε is a stable periodic orbit on $\hat{\mathcal{M}}_\varepsilon$ for the restricted system* (3.2), *then its local unstable manifold, $W^{\mathrm{u}}_{\mathrm{loc}}(O_\varepsilon)$, limits, as $\varepsilon \to 0$, onto the local unstable manifold, $W^{\mathrm{u}}_{\mathrm{loc}}(O_0)$, of the limiting closed curve O_0.*

2. *If O_ε is an unstable periodic orbit on $\hat{\mathcal{M}}_\varepsilon$ for the restricted system* (3.2), *then its local stable manifold, $W^{\mathrm{s}}_{\mathrm{loc}}(O_\varepsilon)$, limits, as $\varepsilon \to 0$, onto the local stable manifold,*

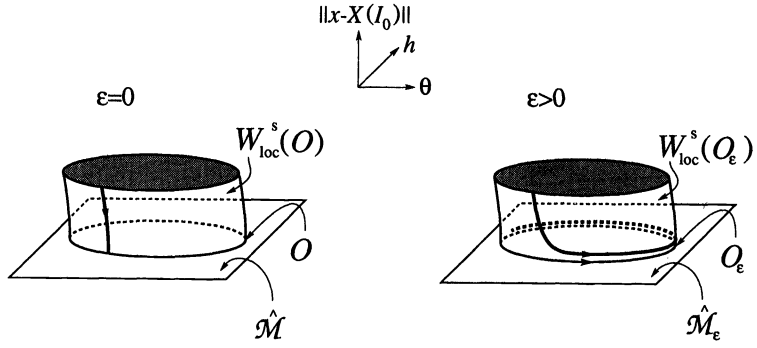


FIG. 8. *As $\varepsilon \to 0$ the local unstable manifold $W^u_{loc}(O_\varepsilon)$ of the periodic orbit O_ε limits onto the local unstable manifold $W^u_{loc}(O_0)$ of the limiting curve O_0. This curve is a periodic orbit of the outer system.*

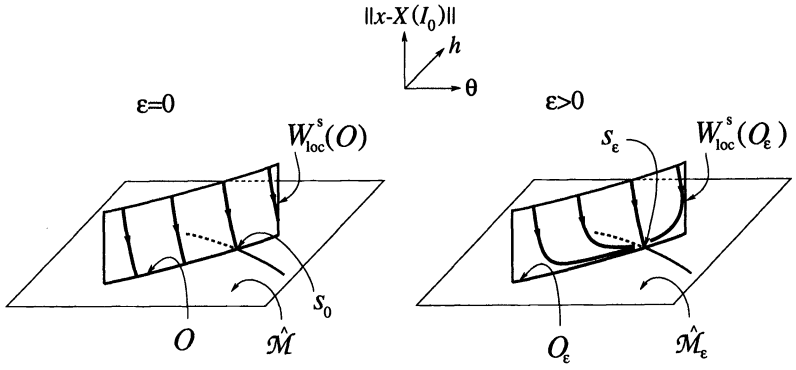


FIG. 9. *When the orbit segment O_ε is the restricted unstable manifold of a saddle s_ε on the annulus $\hat{\mathcal{M}}_\varepsilon$, then its local unstable manifold $W^u_{loc}(O_\varepsilon)$ is a part of the unstable manifold of the saddle s_ε that is all contained in the δ-neighborhood U_δ of the annulus $\hat{\mathcal{M}}$. As $\varepsilon \to 0$, the local unstable manifold $W^u_{loc}(O_\varepsilon)$ limits onto the local unstable manifold $W^u_{loc}(O_0)$ of the limiting curve O_0. This curve is a segment of the unstable manifold of the limiting saddle s_0 for the outer system.*

$W^s_{loc}(O_0)$, *of the limiting closed curve* O_0.

3. *If O_ε is a piece of the restricted stable manifold of a saddle on $\hat{\mathcal{M}}_\varepsilon$ as discussed above, then its local stable manifold, $W^s_{loc}(O_\varepsilon)$, limits, as $\varepsilon \to 0$, onto the local stable manifold, $W^s_{loc}(O_0)$, of the limiting curve O_0.*

4. *If O_ε is a piece of the restricted unstable manifold of a saddle on $\hat{\mathcal{M}}_\varepsilon$ as discussed above, then its local unstable manifold, $W^u_{loc}(O_\varepsilon)$, limits, as $\varepsilon \to 0$, onto the local unstable manifold, $W^u_{loc}(O_0)$, of the limiting curve O_0.*

5. *If c_ε is a sink on the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$ for the restricted system (3.2), then, as $\varepsilon \to 0$, its local unstable manifold limits onto the local unstable manifold of the limiting center, c_0, on the annulus $\hat{\mathcal{M}}$.*

6. *If c_ε is a source on the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$ for the restricted system (3.2), then, as $\varepsilon \to 0$, its local stable manifold limits onto the local stable manifold of the limiting center, c_0, on the annulus $\hat{\mathcal{M}}$.*

In all of the above cases, the local stable and unstable manifolds of the curves O_ε and O_0 and the equilibria c_ε and c_0 are $\mathcal{O}(\sqrt{\varepsilon})$ apart, respectively.

7. Proofs of the main theorems. In this section we finally couple the dynamics near the resonance band in the annulus $\mathcal{M}_\varepsilon$ with the dynamics on the surviving homoclinic orbits using the geometric singular perturbation theory discussed in the previous section.

We recall the limiting homoclinic intersection surface $\Sigma_0^\lambda(\bar\phi, \bar\theta_0)$, which consists of those heteroclinic orbits connecting equilibria on the $h-\theta$ cylinder $\hat{\mathcal{M}}$ that emerge from the $h-\theta$ cylinder $\hat{\mathcal{M}}$ along the line $\theta = \bar\theta_0(I_0,\lambda) - \Delta\theta_-(\bar\phi(I_0,\lambda))$ and return to $\hat{\mathcal{M}}$ along the line $\theta = \bar\theta_0(I_0,\lambda) + \Delta\theta_+(\bar\phi(I_0,\lambda))$. Here, by formulas (3.6),

$$\Delta\theta_+(\phi) = \int_0^\infty \Omega(x^h(s,I_0,\phi),I_0)ds, \qquad \Delta\theta_-(\phi) = \int_{-\infty}^0 \Omega(x^h(s,I_0,\phi),I_0)ds.$$

We now proceed to prove Theorem 1. To do this, we begin by proving two auxiliary propositions. The first proposition is a local transversality result.

PROPOSITION 7.1. *Let $\varepsilon = 0$ and let for $\lambda = \bar\lambda$ the line $\theta = \bar\theta_0 - \Delta\theta_-(\bar\phi)$ intersect transversely the curve $O_{1,0}(\bar\lambda)$, and the line $\theta = \bar\theta_0 + \Delta\theta_+(\bar\phi)$ intersect transversely the curve $O_{2,0}(\bar\lambda)$. Then for all λ near $\bar\lambda$, the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,0}(\lambda))$ of the curve $O_{1,0}(\lambda)$ intersects the limiting homoclinic intersection surface $\Sigma_0^\lambda(\bar\phi,\bar\theta_0)$ transversely inside $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}})$, and the local stable manifold $W^{\mathrm{s}}_{\mathrm{loc}}(O_{2,0}(\lambda))$ of the curve $O_{2,0}(\lambda)$ intersects the limiting homoclinic intersection surface $\Sigma_0^\lambda(\bar\phi,\bar\theta_0)$ transversely inside $W^{\mathrm{s}}_{\mathrm{loc}}(\hat{\mathcal{M}})$.*

Proof. We prove the first part of the proposition; the proof of the second part is almost identical. Recall that the manifold $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}})$ is parametrized by t, h, ϕ and θ_0 in the expression $\big(x^h(t,I_0,\phi), h, \theta^h(t,I_0,\phi)+\theta_0\big)$. The tangent space at any point of $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}})$ is therefore spanned by the vectors

$$\Big(\dot x^h(t,I_0,\phi), 0, \dot\theta^h(t,I_0,\phi)\Big),$$

$$\big(D_\phi x^h(t,I_0,\phi), 0, D_\phi\theta^h(t,I_0,\phi)\big),$$

$$(0,1,0),$$

$$(0,0,1).$$

Now, since the curve $O_{1,0}(\bar\lambda)$ intersects the vertical line $\theta = \bar\theta_0 - \Delta\theta_-(\bar\phi)$ transversely on the annulus $\hat{\mathcal{M}}$, the same must be true for the intersection of the curve $O_{1,0}(\lambda)$ and the vertical line $\theta = \bar\theta_0(I_0,\lambda) - \Delta\theta_-(\bar\phi(I_0,\lambda))$ for all λ close enough to $\lambda = \bar\lambda$. Therefore, the curve $O_{1,0}(\lambda)$ must be expressible as a graph $h = h(\theta)$ near $\theta = \bar\theta_0(I_0,\lambda) - \Delta\theta_-(\bar\phi(I_0,\lambda))$. Thus the manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,0}(\lambda))$ can be parametrized by t, ϕ, and θ_0 in the expression $\big(x^h(t,I_0,\phi), h(\theta^h(-\infty,I_0,\phi)+\theta_0), \theta^h(t,I_0,\phi)+\theta_0\big)$. The tangent space at any point of the manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,0}(\bar\lambda))$ is therefore spanned by the vectors

$$\Big(\dot x^h(t,I_0,\phi), 0, \dot\theta^h(t,I_0,\phi)\Big),$$

$$\left(D_\phi x^h(t,I_0,\phi), \frac{dh}{d\theta}(\theta^h(-\infty,I_0,\phi)+\theta_0)D_\phi\theta^h(-\infty,I_0,\phi), D_\phi\theta^h(t,I_0,\phi)\right),$$

$$\left(0, \frac{dh}{d\theta}(\theta^h(-\infty, I_0, \phi) + \theta_0), 1\right).$$

Finally, the tangent space at any point of the limiting intersection surface $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$ is spanned by the two vectors

$$\left(\dot{x}^h(t, I_0, \phi), 0, \dot{\theta}^h(t, I_0, \phi)\right),$$

$$(0, 1, 0).$$

It is therefore easy to see that the tangent spaces of the manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,0}(\lambda))$ and the limiting intersection surface $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$ add up to the tangent space of the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}})$. □

The second auxiliary proposition uses the first proposition to prove the existence of two special orbits.

PROPOSITION 7.2. *Let the curve $O_{1,\varepsilon}(\lambda)$ be either a stable periodic orbit for the restricted system* (3.2) *on $\hat{\mathcal{M}}_\varepsilon$ or a (restricted) unstable manifold of a saddle for this system. Let the curve $O_{2,\varepsilon}(\lambda)$ be either an unstable periodic orbit for the restricted system* (3.2) *on $\hat{\mathcal{M}}_\varepsilon$ or a (restricted) stable manifold of a saddle for this system. Moreover, let for $\lambda = \bar{\lambda}$ the line $\theta = \bar{\theta}_0 - \Delta\theta_-(\bar{\phi})$ intersect transversely the curve $O_{1,0}(\bar{\lambda})$, and the line $\theta = \bar{\theta}_0 + \Delta\theta_+(\bar{\phi})$ intersect transversely the curve $O_{2,0}(\bar{\lambda})$. Then for all λ near $\lambda = \bar{\lambda}$, and all small enough positive ε, there exists an orbit $a^\lambda_{1,\varepsilon}(t)$ that is contained in both the unstable manifold $W^{\mathrm{u}}(O_{1,\varepsilon}(\lambda))$ and the intersection surface $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$. Likewise, there exists an orbit $a^\lambda_{2,\varepsilon}(t)$ that is contained in both the stable manifold $W^{\mathrm{s}}(O_{2,\varepsilon}(\lambda))$ and the intersection surface $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$. Trajectories on both these orbits are $\mathcal{O}(\sqrt{\varepsilon})$ close (in the $x - h - \theta$ coordinates) to trajectories on the unperturbed counterparts of these orbits for all finite times t.*

Recall the definition of the local unstable manifolds of the curves $O_{1,\varepsilon}(\lambda(\varepsilon))$ and $O_{2,\varepsilon}(\lambda(\varepsilon))$ from §6 in the case when $O_{1,\varepsilon}(\lambda(\varepsilon))$ is the (restricted) unstable manifold of a saddle on the annulus $\hat{\mathcal{M}}_\varepsilon$ or when $O_{2,\varepsilon}(\lambda(\varepsilon))$ is the (restricted) stable manifold of a saddle on $\hat{\mathcal{M}}_\varepsilon$. For an illustration of Proposition 7.2, see Fig. 10.

Proof of Proposition 7.2. As in the proof of the previous proposition, we again show only the first part of this proposition. Let us choose a small enough δ and consider the region $\delta/2 \le \|x - X(I)\| \le \delta$. The previous proposition implies that the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,0}(\lambda))$ and the limiting intersection surface $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$ intersect transversely inside the piece of the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}})$ that satisfies the inequalities $\delta/2 \le \|x - X(I)\| \le \delta$.

Now, for small positive ε, the piece of the homoclinic intersection surface $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ inside $\delta/2 \le \|x - X(I)\| \le \delta$ is $\mathcal{O}(\sqrt{\varepsilon})$ away from the corresponding piece of the surface $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$ and is contained in the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$. By persistence of transverse intersections inside the manifold $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$, the pieces of the surface $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ and the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,\varepsilon}(\lambda))$ that lie inside the region $\delta/2 \le \|x - X(I)\| \le \delta$ still intersect each other. By (local) invariance, this intersection must take place along a segment of an orbit. This orbit segment is $\mathcal{O}(\sqrt{\varepsilon})$ close to the segment of the intersection orbit of the unperturbed local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,0}(\lambda))$ and the limiting intersection surface $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$ that lies in the region $\delta/2 \le \|x - X(I)\| \le \delta$.

The whole orbit, $a^\lambda_{1,\varepsilon}(t)$, which we obtain from this perturbed orbit segment by evolving it in forward and backward time, must be contained in the unstable manifold

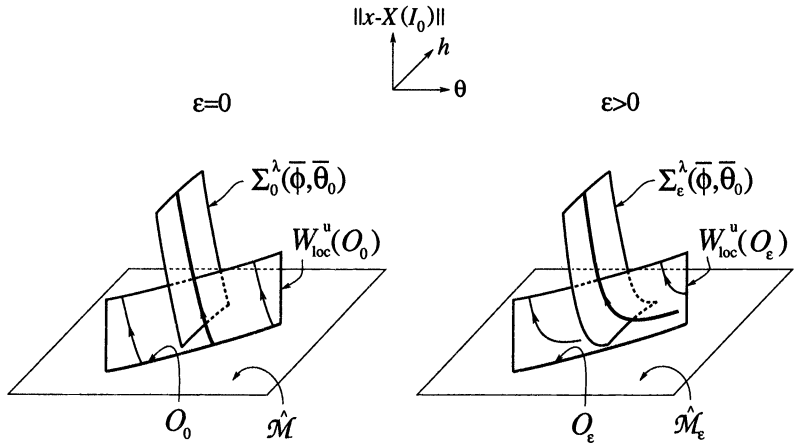


FIG. 10. *The local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,0}(\lambda))$ of the curve $O_{1,0}(\lambda)$ and the limiting homoclinic intersection surface $\Sigma^{\lambda}_{0}(\bar{\phi},\bar{\theta}_0)$ intersect transversely inside the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}})$ of the annulus $\hat{\mathcal{M}}$. Therefore, the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(O_{1,\varepsilon}(\lambda))$ of the orbit segment $O_{1,\varepsilon}(\lambda)$ and the homoclinic intersection surface $\Sigma^{\lambda}_{\varepsilon}(\bar{\phi},\bar{\theta}_0)$ must intersect transversely inside the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}(\hat{\mathcal{M}}_\varepsilon)$ of the annulus $\hat{\mathcal{M}}_\varepsilon$ for small positive ε.*

$W^{\mathrm{u}}(O_{1,\varepsilon}(\lambda))$ by the invariance of this manifold. The $\mathcal{O}(\sqrt{\varepsilon})$ proximity for finite times of the trajectories on this orbit and the intersection orbit of the unperturbed unstable manifold $W^{\mathrm{u}}(O_{1,0}(\lambda))$ and the limiting intersection surface $\Sigma^{\lambda}_{0}(\bar{\phi},\bar{\theta}_0)$ now follows by Proposition 4.2. □

The preceding proposition now renders the following proof.

Proof of Theorem 1. By the previous proposition, there exist two particular orbits on the homoclinic intersection surface $\Sigma^{\lambda}_{\varepsilon}(\bar{\phi},\bar{\theta}_0)$, one forward asymptotic to either the periodic orbit $O_{2,\varepsilon}(\lambda)$ or the saddle $s_{2,\varepsilon}(\lambda)$, and another backward asymptotic to either the periodic orbit $O_{1,\varepsilon}(\lambda)$ or the saddle $s_{1,\varepsilon}(\lambda)$. Also, by the previous proposition and because of equation (3.8), for $\lambda > \bar{\lambda}$, the h-coordinate of any point on one of these orbits is always larger than the h-coordinate of the corresponding point on the other orbit at the same value of θ. For $\lambda < \bar{\lambda}$, the roles are reversed. Therefore, at some $\lambda = \lambda(\varepsilon)$ near $\lambda = \bar{\lambda}$ with $\lambda(0) = \bar{\lambda}$, the two orbits must pass through each other, and thus form a heteroclinic orbit connecting either the periodic orbit $O_{1,\varepsilon}(\lambda(\varepsilon))$ or the saddle $s_{1,\varepsilon}(\lambda(\varepsilon))$ to either the periodic orbit $O_{2,\varepsilon}(\lambda(\varepsilon))$ or the saddle $s_{2,\varepsilon}(\lambda(\varepsilon))$, as claimed. □

An analogous theorem can be proven in the Hamiltonian case. There, two of the three main ingredients of the proof are again Propositions 7.1 and 7.2. However, the transversality condition (3.8) must be dropped, and condition (3.7) is now identical to setting the first component of the Melnikov vector equal to zero. In the proof, an energy argument must be used to show the existence of a heteroclinic connection instead of the transversality argument following from the condition (3.8) used here. Moreover, in the Hamiltonian case, the homoclinic connection will exist for all λ close enough to $\lambda = \bar{\lambda}$. For details, see [40].

To prove Theorem 2, we need to show the existence of the heteroclinic orbit in question, as well as the fact that the manifolds $W^{\mathrm{u}}(O_{1,\varepsilon}(\lambda))$ and $W^{\mathrm{s}}(\hat{\mathcal{M}}_\varepsilon)$ intersect transversely along this heteroclinic orbit. The argument proceeds as follows.

Proof of Theorem 2. Let λ be close to $\lambda = \bar{\lambda}$. Then Proposition 7.2 ensures that

there exists an orbit that is contained in both the unstable manifold $W^u(O_{1,\varepsilon}(\lambda))$ and the intersection surface $\Sigma^\lambda_\varepsilon(\bar\phi,\bar\theta_0)$. Let $a^\lambda_{1,\varepsilon}(t)$ be a trajectory on this orbit, and let $a^\lambda_{1,0}(t)$ be a trajectory on its unperturbed counterpart that starts $\mathcal{O}(\sqrt{\varepsilon})$ away from $a^\lambda_{1,\varepsilon}(t)$ at $t=0$. Let $T>0$ be large enough so that both $a^\lambda_{1,0}(t)$ and $a^\lambda_{1,\varepsilon}(t)$ return inside the neighborhood U_δ at the time $t=T$. By Proposition 4.2, the points $a^\lambda_{1,0}(T)$ and $a^\lambda_{1,\varepsilon}(T)$ are at most $\mathcal{O}(\sqrt{\varepsilon})$ apart. The stable fiber passing through the point $a^\lambda_{1,\varepsilon}(T)$ is $\mathcal{O}(\sqrt{\varepsilon})$ close to the stable fiber passing through the point $a^\lambda_{1,0}(T)$ by Proposition 6.1. Therefore, the base points of these two fibers are $\mathcal{O}(\sqrt{\varepsilon})$ close, as well. But the base point of the fiber that passes through the point $a^\lambda_{1,0}(T)$ is precisely the point $(h,\theta)=\big(0,\bar\theta_0(I_0,\lambda)+\Delta\theta_+(\bar\phi(I_0,\lambda))\big)$. Thus, for all small enough ε, the base point of the fiber through $a^\lambda_{1,\varepsilon}(T)$ must be contained in the basin of attraction $\mathcal{B}_\varepsilon$, which proves the existence part of the theorem.

To show the transversality of the intersection of the unstable manifold $W^u(O_{1,\varepsilon}(\lambda))$ with the stable manifold $W^s(\hat{\mathcal{M}}_\varepsilon)$ of the perturbed annulus $\hat{\mathcal{M}}_\varepsilon$ along the orbit $a_\varepsilon(t)$, we first recall that Proposition 7.1 implies that the intersection of the homoclinic surface $\Sigma^\lambda_\varepsilon(\bar\phi,\bar\theta_0)$ and the unstable manifold $W^u(O_{1,\varepsilon}(\lambda))$ is transverse inside the unstable manifold $W^u(\hat{\mathcal{M}}_\varepsilon)$ of $\hat{\mathcal{M}}_\varepsilon$. We also recall that, by Proposition 4.3, the manifolds $W^s(\hat{\mathcal{M}}_\varepsilon)$ and $W^u(\hat{\mathcal{M}}_\varepsilon)$ intersect along the surface $\Sigma^\lambda_\varepsilon(\bar\phi,\bar\theta_0)$ transversely in the full phase space. This clearly implies that, since $\Sigma^\lambda_\varepsilon(\bar\phi,\bar\theta_0)$ is contained in the stable manifold $W^s(\hat{\mathcal{M}}_\varepsilon)$, this stable manifold and the unstable manifold $W^u(O_{1,\varepsilon}(\lambda))$ must intersect transversely along the heteroclinic orbit $a_\varepsilon(t)$ in the full phase space. Now, either the stable manifold $W^s(c_\varepsilon(\lambda))$ of the equilibrium $c_\varepsilon(\lambda)$ or the stable manifold $W^s(O_{2,\varepsilon}(\lambda))$ of the periodic orbit $O_{2,\varepsilon}(\lambda)$ are neighborhoods of the orbit $a^\lambda_{1,\varepsilon}(T)$ inside the stable manifold $W^s(\hat{\mathcal{M}}_\varepsilon)$ of the annulus $\hat{\mathcal{M}}_\varepsilon$. Therefore, the above transversality argument holds for the manifolds $W^s(c_\varepsilon(\lambda))$ or $W^s(O_{2,\varepsilon}(\lambda))$ in place of the manifold $W^s(\hat{\mathcal{M}}_\varepsilon)$, which concludes our proof. □

Finally, we prove Theorem 3. A different proof of a special case of this theorem with $x\in\mathbb{R}^2$ appeared in [30]. The present proof is included in this paper in order to show how the result of [30] fits in the more general framework that leads at once to all three theorems, and also to extend the proof of [30] to the case when $x\in\mathbb{R}^{2n}$ with $n>1$.

Proof of Theorem 3. First recall that, in the inner limit, the unstable manifold $W^u(\hat{\mathcal{M}})$ of the annulus $\hat{\mathcal{M}}$ is parametrized by t, h, ϕ and θ_0 in the expression

$$\big(x^h(t,I_0,\phi),h,\theta^h(t,I_0,\phi)+\theta_0\big)\,. \tag{7.1}$$

The unstable manifold $W^u(c_0(\lambda))$ of the point $c_0(\lambda)$ is parametrized by t and ϕ in the expression obtained by choosing $h=0$ and $\theta_0=\theta(c_0(\lambda))+\Delta\theta_-(\phi)$ in formula (7.1). Likewise, the limiting homoclinic intersection surface $\Sigma^\lambda_0(\bar\phi,\bar\theta_0)$ is parametrized by t and h in the expression obtained by choosing $\phi=\bar\phi(I_0,\lambda)$ and $\theta_0=\bar\theta_0(I_0,\lambda)$ in formula (7.1).

At $\lambda=\bar\lambda$, the manifold $W^u(c_0(\lambda))$ and the surface $\Sigma^\lambda_0(\bar\phi,\bar\theta_0)$ intersect along a unique orbit given by the expression

$$\big(x^h(t,I_0,\bar\phi(I_0,\bar\lambda)),0,\theta^h(t,I_0,\bar\phi(I_0,\bar\lambda))+\bar\theta_0(I_0,\bar\lambda)\big)\,, \tag{7.2}$$

because, by equation (3.9), we must have $\bar\theta_0(I_0,\bar\lambda)-\Delta\theta_-(\bar\phi(I_0,\bar\lambda))=\theta(c_0(\bar\lambda))$. Moreover, by formula (3.10), the passage of the manifold $W^u(c_0(\lambda))$ and the surface $\Sigma^\lambda_0(\bar\phi,\bar\theta_0)$ through each other along the orbit (7.2) as λ passes through $\lambda=\bar\lambda$ is transverse inside the unstable manifold $W^u(\hat{\mathcal{M}})$.

Recall now the definition of the neighborhood U_δ of the annuli $\hat{\mathcal{M}}$ and $\hat{\mathcal{M}}_\varepsilon$, whose points satisfy the formula $\|x - X(I)\| < \delta$. Outside of a smaller neighborhood, say $U_{\delta/2}$, of the annuli $\hat{\mathcal{M}}$ and $\hat{\mathcal{M}}_\varepsilon$, the homoclinic intersection surfaces $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ and $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$ are $\mathcal{O}(\sqrt{\varepsilon})$ close to each other. Moreover, the local unstable manifold $W^{\rm u}_{\rm loc}(c_\varepsilon(\lambda))$ varies smoothly with λ and $\sqrt{\varepsilon}$ down to and including $\varepsilon = 0$ inside the neighborhood U_δ by Proposition 6.3. Hence, it follows from the discussion in the previous paragraph that, in the region $\delta/2 < \|x - X(I)\| < \delta$, the mainfold $W^{\rm u}_{\rm loc}(c_\varepsilon(\lambda))$ and the homoclinic intersection surface $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ must pass through each other transversely inside the local unstable manifold $W^{\rm u}_{\rm loc}(\hat{\mathcal{M}}_\varepsilon)$ as λ varies through some $\lambda = \lambda(\varepsilon)$. The function $\lambda(\varepsilon)$ varies smoothly with $\sqrt{\varepsilon}$, and its value at $\varepsilon = 0$ is $\bar{\lambda}$. Moreover, the intersection of $W^{\rm u}_{\rm loc}(c_\varepsilon(\lambda))$ and $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ at $\lambda = \lambda(\varepsilon)$ inside the region $\delta/2 < \|x - X(I)\| < \delta$ takes place along an orbit segment that is $\mathcal{O}(\sqrt{\varepsilon})$ close to an appropriate segment of the orbit (7.2).

Let $a_0(t)$ and $a_\varepsilon(t)$ be two trajectories that start $\mathcal{O}(\sqrt{\varepsilon})$ away from each other at $t = 0$, and lie on the intersections of the unperturbed and the perturbed unstable manifolds, $W^{\rm u}(c_0(\bar{\lambda}))$ and $W^{\rm u}(c_\varepsilon(\lambda(\varepsilon)))$ with the homoclinic surfaces $\Sigma_0^\lambda(\bar{\phi}, \bar{\theta}_0)$ and $\Sigma_\varepsilon^\lambda(\bar{\phi}, \bar{\theta}_0)$ at $\lambda = \bar{\lambda}$ and $\lambda = \lambda(\varepsilon)$, respectively. Since $\lambda(\varepsilon)$ and $\bar{\lambda}$ are at most $\mathcal{O}(\sqrt{\varepsilon})$ apart for small ε, we can proceed as in the proof of Theorem 2 to show that these trajectories are $\mathcal{O}(\sqrt{\varepsilon})$ close to each other for all times up to and including some large enough $T > 0$, and that the stable fibers passing through the respective points $a_\varepsilon(T)$ and $a_0(T)$ are also at most $\mathcal{O}(\sqrt{\varepsilon})$ apart. Thus, as in the proof of Theorem 2, we conclude that the trajectory $a_\varepsilon(t)$ is attracted to the same object as the points in the set $\mathcal{B}_\varepsilon$, which proves the theorem. □

We now make a remark about the uniqueness of the heteroclinic orbits discussed in Theorems 1–3. This remark is in place because the families of stable and unstable fibers described in Proposition 6.1 that foliate the manifolds $W^{\rm s}(\hat{\mathcal{M}}_\varepsilon)$ and $W^{\rm u}(\hat{\mathcal{M}}_\varepsilon)$ need not be unique; see [37]. However, for $\varepsilon > 0$, the equilibria and periodic orbits that the heteroclinic orbits in question connect are by assumption hyperbolic and, thus, unique. Their stable and unstable manifolds are therefore also unique, and so must be the heteroclinic orbits that arise as the intersections of these manifolds.

8. An example. We consider a four-parameter family of problems in which a Duffing oscillator is coupled to an anharmonic oscillator, described by the system of equations

$$\dot{p} = \mu^2 q(I - q^2) - \varepsilon\alpha p, \tag{8.1a}$$

$$\dot{q} = p, \tag{8.1b}$$

$$\dot{I} = -\varepsilon I \sin\theta - \varepsilon\beta I - \varepsilon\gamma p^2, \tag{8.1c}$$

$$\dot{\theta} = I - 1 - \frac{1}{2}\mu^2 q^2 - \varepsilon\cos\theta, \tag{8.1d}$$

where μ, α, β, and γ are positive parameters, and $\varepsilon \ll 1$ is a small positive parameter. This system is of the form

$$\dot{p} = -\frac{\partial H(p,q,I,\theta)}{\partial q} - \varepsilon\alpha p, \qquad \dot{q} = \frac{\partial H(p,q,I,\theta)}{\partial p},$$

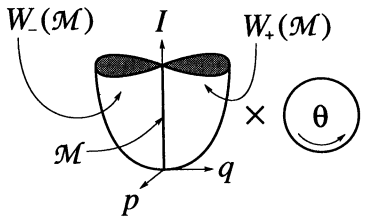


FIG. 11. *The unperturbed invariant annulus, $\mathcal{M}$, and its homoclinic manifolds, $W_+(\mathcal{M})$ and $W_-(\mathcal{M})$ for the Duffing oscillator-harmonic oscillator example.*

$$\dot{I} = -\frac{\partial H(p,q,I,\theta)}{\partial \theta} - \varepsilon\beta I - \varepsilon\gamma p^2, \qquad \dot{\theta} = \frac{\partial H(p,q,I,\theta)}{\partial I},$$

where

$$\begin{aligned} H(p,q,I,\theta) &= H_0(p,q,I) + \varepsilon H_1(p,q,I,\theta) \\ &= \frac{1}{2}I^2 - I + \frac{1}{2}p^2 - \frac{1}{2}\mu^2 q^2\left(I - \frac{1}{2}q^2\right) - \varepsilon I\cos\theta. \end{aligned} \tag{8.2}$$

It is easy to see that equations (8.1) fall into the same category as (2.1).

The unperturbed equations corresponding to (8.1) are

$$\dot{p} = \mu^2 q(I - q^2), \tag{8.3a}$$

$$\dot{q} = p, \tag{8.3b}$$

$$\dot{I} = 0, \tag{8.3c}$$

$$\dot{\theta} = I - 1 - \frac{1}{2}\mu^2 q^2, \tag{8.3d}$$

which can be derived from the unperturbed Hamiltonian

$$H_0(p,q,I) = \frac{1}{2}I^2 - I + \frac{1}{2}p^2 - \frac{1}{2}\mu^2 q^2\left(I - \frac{1}{2}q^2\right). \tag{8.4}$$

The unstable invariant annulus $\mathcal{M}$ is located at $(p,q) = (0,0)$, and can be bounded by any I_1 and I_2, with $0 < I_1 < 1 < I_2$. It is foliated by periodic orbits $p = q = 0$, $I =$ constant, and $\theta = (I-1)t + \theta_0$. The orbit at $I = 1$ is a circle of equilibria, and clearly, the frequency $I - 1$ passes through zero transversely there, so that the resonance Assumption 3 is satisfied. The annulus $\mathcal{M}$ is connected to itself by a pair of three-dimensional homoclinic manifolds, $W_+(\mathcal{M})$ and $W_-(\mathcal{M})$, as is shown in Fig. 11. The manifolds $W_+(\mathcal{M})$, and $W_-(\mathcal{M})$ are parametrized by t, I, and θ_0 in the homoclinic solutions

$$p = p^h(t,I) = \mp\sqrt{2}\mu I\,\mathrm{sech}(\mu\sqrt{I}t)\tanh(\mu\sqrt{I}t), \tag{8.5a}$$

$$q = q^h(t,I) = \pm\sqrt{2I}\,\mathrm{sech}(\mu\sqrt{I}t), \tag{8.5b}$$

$$I = I, \tag{8.5c}$$

$$\theta = \theta^h(t, I) + \theta_0 = (I-1)t - \mu\sqrt{I}\tanh(\mu\sqrt{I}t) + \theta_0. \tag{8.5d}$$

(In this example, the number of dimensions is too small to require the additional parameters ϕ.) From (8.5d) we find the angle difference $\Delta\theta$ between the end points of any heteroclinic orbit connecting pairs of equilibria on $\mathcal{M}$ at $I = 1$ to be $\Delta\theta = -2\mu$.

Even for nonzero ε, the set $p = q = 0$ is invariant. Thus, we can take the perturbed annulus $\mathcal{M}_\varepsilon$ to be the same as the annulus $\mathcal{M}$. At the resonance, $I = 1$, the Melnikov function, $M(I, \theta_0, \alpha, \beta, \gamma)$, can be computed explicitly [49]. This is because the integrand in formula (2.6) in this case reduces to

$$-\frac{dH_1}{dt}(p, q, 1, \theta) - \alpha p^2 + \frac{1}{2}\mu^2\beta q^2 + \frac{1}{2}\mu^2\gamma p^2 q^2,$$

where $p = p^h(t, 1)$, $q = q^h(t, 1)$, and $\theta = \theta^h(t, 1) + \theta_0$. Thus, the Melnikov function becomes

$$\begin{aligned} M(1, \theta_0, \alpha, \beta, \gamma) &= H_1(0, 0, 1, \theta_0 + \mu) - H_1(0, 0, 1, \theta_0 - \mu) - \frac{4}{3}\alpha\mu + 2\beta\mu + \frac{8}{15}\gamma\mu^3 \\ &= -\cos(\theta_0 + \mu) + \cos(\theta_0 - \mu) - \frac{4}{3}\alpha\mu + 2\beta\mu + \frac{8}{15}\gamma\mu^3 \\ &= 2\sin\mu\sin\theta_0 - \frac{4}{3}\alpha\mu + 2\beta\mu + \frac{8}{15}\gamma\mu^3 \end{aligned} \tag{8.6}$$

and is the same on both homoclinic manifolds, $W_+(\mathcal{M})$ and $W_-(\mathcal{M})$.

When μ is not a multiple of π, this Melnikov function has *transverse* zeros in θ_0 at $\theta_0 = \bar{\theta}_{0,1}$ and $\theta_0 = \bar{\theta}_{0,2} = \pi - \bar{\theta}_{0,1}$, provided that

$$\left|\frac{\mu}{\sin\mu}\left(\frac{2}{3}\alpha - \beta - \frac{4}{15}\gamma\mu^2\right)\right| < 1. \tag{8.7}$$

For all admissible α, β, and γ, the stable and unstable manifolds $W^s(\mathcal{M}_\varepsilon)$ and $W^u(\mathcal{M}_\varepsilon)$ intersect transversely along two symmetric pairs of two-dimensional homoclinic surfaces, $\Sigma^{\alpha,\beta,\gamma}_{\pm,\varepsilon}(\bar{\theta}_{0,1})$ and $\Sigma^{\alpha,\beta,\gamma}_{\pm,\varepsilon}(\bar{\theta}_{0,2})$.

The restricted system, (3.2), at the resonance at $I = 1$ for this example is

$$h' = -(1 + \sqrt{\varepsilon}h)\sin\theta - \beta(1 + \sqrt{\varepsilon}h), \qquad \theta' = h - \sqrt{\varepsilon}\cos\theta, \tag{8.8}$$

and the limiting outer system is

$$h' = -\sin\theta - \beta, \qquad \theta' = h. \tag{8.9}$$

The rescaled Hamiltonian is

$$\mathcal{H}(h, \theta) = \frac{1}{2}h^2 - \cos\theta + \beta\theta, \tag{8.10}$$

which is the Hamiltonian of the pendulum subjected to a constant torque.

The phase portrait of the rescaled $h - \theta$ phase cylinder $\mathcal{M}_0$ of the equations (8.9) is shown in Fig. 12. There are two equilibria on this phase cylinder, a center, c_0, at $(h, \theta) = (0, -\arcsin\beta)$, and a saddle, s_0, at $(h, \theta) = (0, -\pi + \arcsin\beta)$. The two

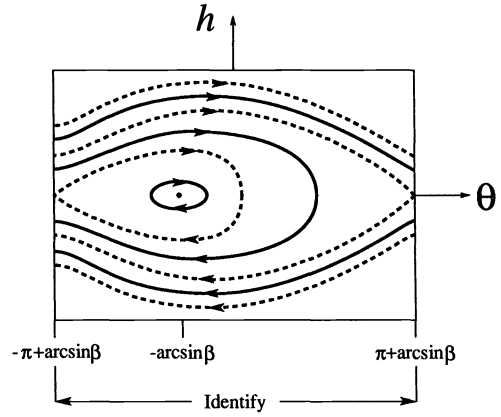


FIG. 12. *The phase portrait of the $h - \theta$ cylinder $\hat{\mathcal{M}}$ at the resonance. Broken lines represent the stable and unstable manifolds, $\mathcal{W}^{s}(s_0)$ and $\mathcal{W}^{u}(s_0)$, of the saddle s_0.*

branches of the stable and unstable manifolds, $\mathcal{W}^{s}(s_0)$ and $\mathcal{W}^{u}(s_0)$, to the right of the saddle s_0 coincide to form a separatrix that encloses a family of periodic orbits nested around the center. The two branches of the manifolds $\mathcal{W}^{s}(s_0)$ and $\mathcal{W}^{u}(s_0)$ to the left of the saddle s_0 wind around the cylinder $\hat{\mathcal{M}}$ toward $h = +\infty$ and $h = -\infty$, respectively. For small positive $\sqrt{\varepsilon}$, the saddle s_0 persists as a saddle, s_ε, the center c_0 becomes a sink, c_ε, and the separatrix breaks. The top branch of the unstable manifold, $\mathcal{W}^{u}(s_\varepsilon)$, of the perturbed saddle s_ε falls into the sink c_ε. No periodic orbits are left in this system, and all the points that lie in any compact domain that is all contained inside the unperturbed separatrix asymptote to the sink c_ε.

The inner system is

$$\dot{p} = \mu^2 q(1 - q^2), \tag{8.11a}$$

$$\dot{q} = p, \tag{8.11b}$$

$$\dot{h} = 0, \tag{8.11c}$$

$$\dot{\theta} = -\frac{1}{2}\mu^2 q^2. \tag{8.11d}$$

In the phase space of this system, the two symmetric pairs of homoclinic intersection surfaces, $\Sigma_{\pm,\varepsilon}^{\alpha,\beta,\gamma}(\bar{\theta}_{0,1})$ and $\Sigma_{\pm,\varepsilon}^{\alpha,\beta,\gamma}(\bar{\theta}_{0,2})$ (when the inequality (8.7) shows that they exist), collapse smoothly onto the pairs of surfaces, $\Sigma_{\pm,0}^{\alpha,\beta,\gamma}(\bar{\theta}_{0,1})$ and $\Sigma_{\pm,0}^{\alpha,\beta,\gamma}(\bar{\theta}_{0,2})$, parametrized by the expressions (8.5) with $I = 1$, $\theta_0 = \bar{\theta}_{0,1}$ or $\bar{\theta}_{0,2}$, and arbitrary h.

We now demonstrate that this example satisfies the conditions of Theorems 1–3. To do so, we consider the case when $\mu \ll \beta < 1$. We assume that $\tilde{\gamma} = \mu^2\gamma = \mathcal{O}(1)$, and let α play the role of the parameter λ. We will show that, for appropriately chosen α and $\tilde{\gamma}$, orbits homoclinic to the saddle s_ε, orbits connecting s_ε to the sink c_ε, and orbits homoclinic to the sink c_ε exist. In fact, due to the symmetry of the problem, all such orbits always occur in pairs: one on the surface $\Sigma_{+,\varepsilon}^{\alpha,\beta,\gamma}(\bar{\theta}_{0,1})$ and the other one on $\Sigma_{-,\varepsilon}^{\alpha,\beta,\gamma}(\bar{\theta}_{0,2})$.

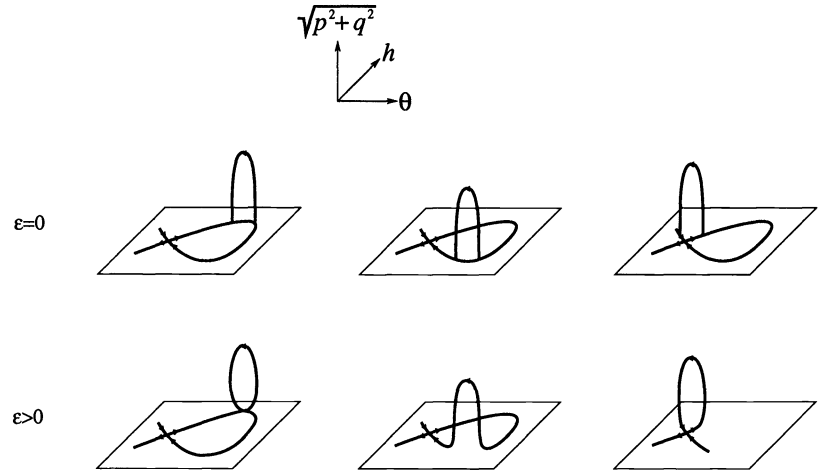


FIG. 13. *The three types of orbits homoclinic to the saddle s_ε, whose existence follows from Theorem* 1.

First, we use Theorem 1 to find pairs of orbits homoclinic to the saddle s_ε. In fact, three different types of such homoclinic orbits exist. They are shown in Fig. 13, and to prove their existence, we proceed as follows. From (8.6), we recall that the equation for any zero, $\bar{\theta}_0$, of the Melnikov function satisfies the equation

$$2\sin\mu\sin\bar{\theta}_0 - \frac{4}{3}\alpha\mu + 2\beta\mu + \frac{8}{15}\tilde{\gamma}\mu = 0.$$

Moreover, formula (3.7) for this example reads

$$\begin{aligned}\frac{1}{2}h^2(\bar{\theta}_0+\mu) - \frac{1}{2}h^2(\bar{\theta}_0-\mu) &= \cos(\bar{\theta}_0+\mu) - \beta\mu - \cos(\bar{\theta}_0-\mu) - \beta\mu \\ &= -2\sin\mu\sin\bar{\theta}_0 - 2\beta\mu \\ &= -\frac{4}{3}\alpha\mu + \frac{8}{15}\tilde{\gamma}\mu \\ &= 0.\end{aligned}$$

Thus, in order to find an orbit homoclinic to the point s_ε, we must solve the equations

$$\sin\bar{\theta}_0 = -\beta\frac{\mu}{\sin\mu} = -\beta + \mathcal{O}(\mu^2),$$

and

$$5\alpha - 2\tilde{\gamma} = 0.$$

We thus obtain

$$\bar{\theta}_{0,1} = -\arcsin\beta + \mathcal{O}(\mu^2)$$

and

$$\bar{\theta}_{0,2} = -\pi + \arcsin\beta + \mathcal{O}(\mu^2).$$

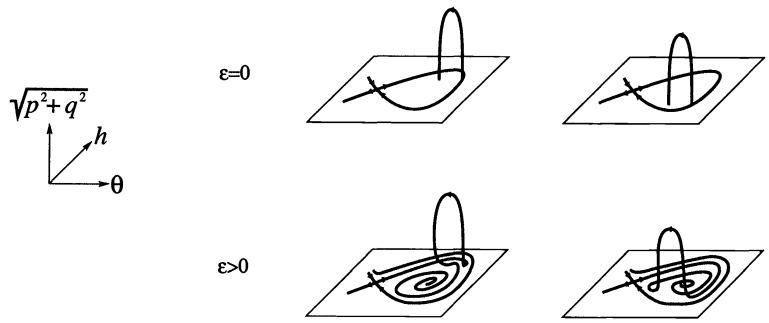


FIG. 14. *The two types of heteroclinic orbits connecting the saddle s_ε and the spiral-saddle c_ε, whose existence follows from Theorem 2.*

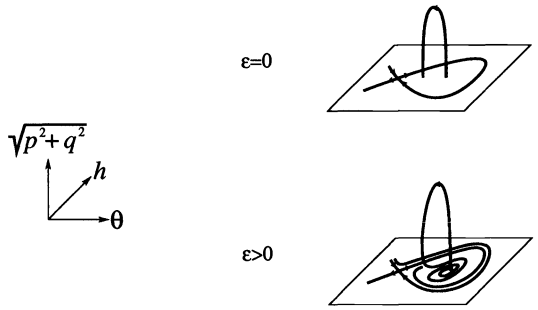


FIG. 15. *An orbit homoclinic to the spiral-saddle c_ε, whose existence follows from Theorem 3.*

This implies that the line $\theta = \bar{\theta}_{0,1}$ passes an $\mathcal{O}(\mu^2)$ distance away from the center c_0, and that the line $\theta = \bar{\theta}_{0,2}$ passes an $\mathcal{O}(\mu^2)$ distance away from the saddle s_0, so that the desired intersections exist and are at the same height, h, whenever $\alpha = 2\tilde{\gamma}/5$.

It can be shown that, in general, if $\mu = n\pi + \delta$, with some nonnegative integer n and $|\delta| \ll \beta < 1$, we obtain orbits homoclinic to the saddle s_ε that wind n times around the cylinder $\mathcal{M}_\varepsilon$ before returning to it. (See [35].)

If we choose α and $\tilde{\gamma}$ so that $5\alpha - 2\tilde{\gamma} > 0$ but sufficiently small, then the hypotheses of Theorem 2 are satisfied, and there exist two pairs of connections between the saddle s_ε and the sink c_ε, as shown in Fig. 14.

Finally, we show the existence of orbits homoclinic to the point c_ε, shown in Fig. 15. One condition for a pair of such orbits to exist is that the line $\theta = \bar{\theta}_0 + \mu$ must pass through the unperturbed equilibrium c_0. This happens when $\cos(\bar{\theta}_0 + \mu) = \sqrt{1-\beta^2}$ and $\sin(\bar{\theta}_0 + \mu) = -\beta$, hence formula (8.6) implies the equation

$$-\sqrt{1-\beta^2} + \sqrt{1-\beta^2}\cos 2\mu - \beta \sin 2\mu - \frac{4}{3}\alpha\mu + 2\beta\mu + \frac{8}{15}\tilde{\gamma}\mu = 0,$$

which, when $\mu \ll \beta < 1$, yields

$$\alpha = \frac{2}{5}\tilde{\gamma} - \frac{3}{2}\mu\sqrt{1-\beta^2} + \mathcal{O}(\mu^2).$$

The second condition is that the point $(h, \theta) = (0, \bar{\theta}_0 - \mu)$ be contained inside the separatrix that encircles the equilibrium point at $(h, \theta) = (0, \bar{\theta}_0 + \mu)$, which is clearly satisfied for $\mu \ll \beta < 1$. When both of these conditions are satisfied, it follows from

Theorem 3 that a pair of orbits homoclinic to the equilibrium c_ε exists. These orbits are of the so-called Šilnikov type [50]; the chaotic dynamics created by such a pair are discussed, for instance in [30].

In conclusion, even this simple example shows the richness of the various homoclinic orbits that may emerge under perturbation from orbits homoclinic to an unstable circle of equilibria that breaks up into a resonance band. Furthermore, this example also shows the ease with which Theorems 1–3 can be applied to specific situations, and thus reveals the potential power of the method for finding orbits homoclinic to resonance bands in solving physical and engineering problems.

Acknowledgments. The author is grateful to A. Calini, R. Camassa, N. Ercolani, Z. Feng, D. Hobson, D. D. Holm, T. Kaper, D. McLaughlin, M. Levi, C. C. Lim, V. Roytburd, S. Wiggins, and C. Xiong for stimulating conversations and to the Theoretical Division and the Center for Nonlinear Studies at the Los Alamos National Laboratory for their hospitality and support during the summers of 1992 through 1995.

REFERENCES

[1] H. POINCARÉ, *Les Méthodes Nouvelles de la Mécanique Celeste*, Gauthier-Villars, Paris, 1899.
[2] V. K. MELNIKOV, *On the stability of the center for time periodic perturbations*, Trans. Moscow Math., 12 (1963), pp. 1–57.
[3] P. J. HOLMES, *A nonlinear oscillator with a strange attractor*, Philos. Trans. Roy. Soc. London, Ser. A, 292 (1979), pp. 419–448.
[4] S. N. CHOW, J. HALE, AND J. MALLET-PARET, *An example of bifurcation to homoclinic orbits*, J. Differential Equations, 37 (1980), pp. 351–373.
[5] B. D. GREENSPAN AND P. J. HOLMES, *Repeated resonance and homoclinic bifurcation in a periodically forced family of oscillators*, SIAM J. Math. Anal., 15 (1984), pp. 69–97.
[6] V. I. ARNOLD, *Instability of dynamical systems with many degrees of freedom*, Soviet Math. Dokl., 5 (1964), pp. 581–585.
[7] P. J. HOLMES AND J. E. MARSDEN, *Horseshoes in perturbation of Hamiltonian systems with two degrees of freedom*, Comm. Math. Phys., 82 (1982), pp. 523–544.
[8] ———, *Melnikov's method and Arnold diffusion for perturbations of integrable Hamiltonian systems*, J. Math. Phys., 23 (1982), pp. 669–675.
[9] C. ROBINSON, *Sustained resonance for a nonlinear system with slowly varying coefficients*, SIAM J. Math. Anal., 14 (1983), pp. 847–860.
[10] L. M. LERMAN AND IA. L. UMANSKI, *On the existence of separatrix loops in four-dimensional systems similar to integrable Hamiltonian systems*, Prikl. Math. Mech. U.S.S.R., 47 (1984), pp. 335–340.
[11] K. J. PALMER, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225–256.
[12] J. GRUENDLER, *The existence of homoclinic orbits and the method of Melnikov for systems in* $\mathbb{R}^n$, SIAM J. Math. Anal., 16 (1985), pp. 907–931.
[13] P. J. HOLMES, *Chaotic motions in a weakly nonlinear model for surface waves*, J. Fluid Mech., 162 (1986), pp. 365–388.
[14] S. WIGGINS AND P. J. HOLMES, *Homoclinic orbits in slowly varying oscillators*, SIAM J. Math. Anal., 18 (1987), pp. 612–629.
[15] C. ROBINSON, *Horseshoes for autonomous Hamiltonian systems using the Melnikov integral*, Ergodic Theory Dynamical Systems, 8* (1988), pp. 395–409.
[16] S. W. SHAW AND S. WIGGINS, *Chaotic dynamics of a whirling pendulum*, Physica D, 31 (1988), pp. 190–211.
[17] S. WIGGINS, *Global Bifurcations and Chaos: Analytical Methods*, Springer-Verlag, New York, 1988.
[18] A. N. KOLMOGOROV, *On conservation of conditionally periodic motion for a small change in Hamilton's function*, Dokl. Akad. Nauk. UzSSR, 98:4 (1954), pp. 525–530.
[19] V. I. ARNOLD, *Proof of a theorem of A. N. Kolmogorov on the invariance of quasiperiodic motions under small perturbations of the Hamiltonian*, Uspekhi Mat. Nauk, 18 (1965), pp. 13–41.
[20] J. MOSER, *A rapidly converging iteration method and nonlinear partial differential equations*,

I, Ann. Scuola Norm. Sup. Pisa, Cl. Sci. (3), 20 (1966), pp. 265–315.
[21] J. MOSER, *A rapidly converging iteration method and nonlinear partial differential equations*, II, Ann. Scuola Norm. Sup. Pisa, Cl. Sci. (3), 20 (1966), pp. 499–535.
[22] S. M. GRAFF, *On the conservation of hyperbolic invariant tori for Hamiltonian systems*, J. Differential Equations, 15 (1974), pp. 1–69.
[23] Z. C. FENG AND P. R. SETHNA, *Symmetry-breaking bifurcations in resonant surface waves*, J. Fluid. Mech., 199 (1989), pp. 495–518.
[24] ———, *Global bifurcation and chaos in parametrically forced systems with one-one resonance*, Dynamics Stability Systems, 5, (1990), pp. 201–225.
[25] D. DAVID, D. D. HOLM, AND M. V. TRATNIK, *Hamiltonian chaos in nonlinear optical polarization dynamics*, Phys. Rep., 187 (1990), pp. 283–367.
[26] D. D. HOLM, G. KOVAČIČ, AND B. SUNDARAM, *Chaotic laser-matter interaction*, Phys. Lett. A, 154 (1991), pp. 346–352.
[27] X. L. YANG AND P. R. SETHNA, *Local and global bifurcations in parametrically excited vibrations of nearly square plates*, Internat. J. Non-linear Mech., 26 (1991), pp. 199–220.
[28] D. D. HOLM AND G. KOVAČIČ, *Homoclinic chaos in a laser-matter system*, Physica D, 56 (1992), pp. 270–300.
[29] A. ACEVES, D. D. HOLM, G. KOVAČIČ, *Chaotic dynamics due to competition among degenerate modes in a ring cavity laser*, Phys. Lett. A, 161 (1992), pp. 499–505.
[30] G. KOVAČIČ AND S. WIGGINS, *Orbits homoclinic to resonances with an application to chaos in a model of the forced and damped sine-Gordon equation*, Physica D, 57 (1992), pp. 185–225.
[31] Z. C. FENG AND S. WIGGINS, *On the existence of chaos in a class of two-degree-of-freedom, damped, parametrically forced mechanical systems with broken $O(2)$ symmetry*, Z. Angew. Math. Phys., 44 (1992), pp. 201–248.
[32] J. GUCKENHEIMER AND P. J. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.
[33] S. WIGGINS, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, 1990.
[34] V. I. ARNOLD (ed.), *Dynamical Systems III*, in Encyclopedia of Mathematical Sciences, Vol. 3, Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, 1988.
[35] G. KOVAČIČ, *Dissipative dynamics of orbits homoclinic to a resonance band*, Phys. Lett. A, 167 (1992), pp. 143–150.
[36] D. W. MCLAUGHLIN, E. A. OVERMAN, S. WIGGINS, AND C. XIONG, *Homoclinic behavior for a 2 mode truncation of NLS*, 1993, preprint.
[37] N. FENICHEL, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.
[38] K. SAKAMOTO, *Invariant manifolds in singular perturbation problems for ordinary differential equations*, Proc. Roy. Soc. Edinburgh, Sect. A, 116 (1990), pp. 45–78.
[39] G. KOVAČIČ, *Hamiltonian dynamics of orbits homoclinic to a resonance band*, Phys. Lett. A, 167 (1992), 137–142.
[40] ———, *Singular perturbation theory for homoclinic orbits in a class of near-integrable Hamiltonian systems*, Dynamics Differential Equations, 5 (1993), pp. 559–597.
[41] G. HALLER AND S. WIGGINS, *Orbits homoclinic to resonances: the Hamiltonian case*, Physica D, 66 (1993), pp. 298–346.
[42] A. R. BISHOP, R. FLESCH, M. G. FOREST, D. W. MCLAUGHLIN, AND E. A. OVERMAN, *Correlations between chaos in a perturbed sine-Gordon equation and a truncated model system*, SIAM J. Math. Anal., 21 (1990), pp. 1511–1536.
[43] X. M. GU AND P. R. SETHNA, *Resonant surface waves and chaotic phenomena*, J. Fluid Mech., 183 (1987), pp. 543–565.
[44] N. FENICHEL, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–225.
[45] ———, *Asymptotic stability with rate conditions*, Indiana Univ. Math. J., 23 (1974), pp. 1109–1137.
[46] ———, *Asymptotic stability with rate conditions*, II, Indiana Univ. Math. J., 26 (1977), pp. 81–93.
[47] V. I. ARNOLD, *Ordinary Differential Equations*, MIT Press, Cambridge, MA., 1973.
[48] C. K. R. T. JONES AND N. KOPELL, *Tracking invariant manifolds with differential forms in singularly perturbed systems*, J. Differential Equations, 108 (1994), pp. 64–88.
[49] Z. C. FENG, private communication, 1990.
[50] L. P. ŠILNIKOV, *A case of the existence of a denumerable set of periodic motions*, Soviet Math. Dokl., 6 (1965), pp. 163–166.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1644–1646, November 1995

014

A SIMPLE PROOF OF FRYANT'S THEOREM*

M. K. VEMURI†

Abstract. In [A. Fryant, *SIAM J. Math. Anal.*, 22 (1991), pp. 268–271.], the spherical harmonics of degree less than or equal to k in $\mathbb{R}^{n-1}$ were used to generate the spherical harmonics of degree k in $\mathbb{R}^n$. Invariant theory was used to show that the resulting set of spherical harmonics is orthogonal. A simple calculation for accomplishing this directly is given here.

Key words. spherical harmonics, irreducible representation

AMS subject classifications. 33A45, 31B99

Let Σ_{n-1} denote the unit sphere in $\mathbb{R}^n$ and $\langle f, g\rangle$ be the usual inner product on Σ_{n-1}. That is,

$$\langle f, g\rangle = \int_{\Sigma_{n-1}} f(x)\overline{g(x)}dx.$$

The following theorem was proved in [1].

THEOREM 1. *Let $P_1, \ldots, P_{d_n^k}$ be orthonormal spherical harmonics in $\mathbb{R}^{n-1}$ of degree less than or equal to k. If $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $t = (t_1, \ldots, t_{n-1}) \in \Sigma_{n-2}$, then*

(1) $(x_1 + it_1x_2 + \cdots + it_{n-1}x_n)^k = \sum_{j=1}^{d_n^k} Y_k^j(x)P_j(t)$, *where $Y_k^j, j = 1, 2, \ldots, d_n^k$, are homogeneous harmonic polynomials of degree k in $\mathbb{R}^n$, and*

(2) $\langle Y_k^j, Y_k^l\rangle = 0$ *if $j \neq l$.*

In [1] invariant theory was used to prove the second assertion. We give a simpler proof which avoids invariant theory. Let H_n^k denote the space of harmonic homogeneous polynomials of degree k in $\mathbb{R}^n$. For $f, g \in H_n^k$, let

$$(f, g) = f\left(\frac{\partial}{\partial x}\right)\overline{g},$$

that is, the differential operator $f\left(\frac{\partial}{\partial x}\right)$ acting on the function $\overline{g}$. The result is a constant. It is easily seen that $(\cdot, \cdot)$ is an inner product on H_n^k.

LEMMA 1. *There is a constant c_n^k such that for all $f, g \in H_n^k$,*

$$(f, g) = c_n^k\langle f, g\rangle.$$

Proof. Recall [3] that H_n^k is an irreducible representation of $O(n)$ under the action

$$(Af)(x) = f(A^{-1}x),$$

where $A \in O(n)$, $f \in H_n^k$, and $x \in \mathbb{R}^n$. It is easily seen that both $\langle\cdot, \cdot\rangle$ and $(\cdot, \cdot)$ are invariant under this action.

Let K be the compact set $\{f|\langle f, f\rangle = 1\}$. Let $f_0 \in K$ be the point at which the continuous function $f \mapsto (f, f)$ achieves its maximum c_n^k on K. Thus $c_n^k\langle\cdot, \cdot\rangle - (\cdot, \cdot)$ is positive semidefinite, and $c_n^k\langle f_0, f_0\rangle - (f_0, f_0) = 0$. So $c_n^k\langle\cdot, \cdot\rangle - (\cdot, \cdot)$ is degenerate. Let $N \neq \{0\}$ be its kernel. Then N is $O(n)$ invariant. Since H_n^k is irreducible, $N = H_n^k$, i.e., $c_n^k\langle\cdot, \cdot\rangle - (\cdot, \cdot) \equiv 0$. It follows that $(\cdot, \cdot) = c_n^k\langle\cdot, \cdot\rangle$, and the theorem is proved.

* Received by the editors November 3, 1993; accepted for publication (in revised form) March 23, 1994.

† University of Chicago, Chicago, Illinois 60637.

We now show that $\{Y_k^j\}_{j=1}^{d_n^k}$ is an orthogonal set with respect to the inner product $(\cdot,\cdot)$. Orthogonality with respect to $\langle\cdot,\cdot\rangle$ will follow by the previous lemma.

LEMMA 2. *Let $f \in C^\infty(\mathbb{R}^n)$ and X, a first order homogeneous linear differential operator such that $X^2 f = 0$. Then*

$$X^k(f^k) = k!(Xf)^k.$$

Proof. Note that for all j, $X(Xf)^j = j(Xf)^{j-1}X^2 f = 0$. Therefore

$$\begin{aligned} X^k f^k &= X^{k-1}Xf^k = X^{k-1}(kf^{k-1}Xf) = X^{k-2}X(kf^{k-1}Xf) \\ &= X^{k-2}(k(k-1)f^{k-2}(Xf)^2) \\ &= \cdots = k!(Xf)^k. \end{aligned}$$

THEOREM 2. $(Y_k^j, Y_k^l) = 0$ *if* $j \neq l$.

Proof.

$$\begin{aligned} (1)\quad &((x_1 + it_1x_2 + \cdots + it_{n-1}x_n)^k, (x_1 + is_1x_2 + \cdots + is_{n-1}x_n)^k) \\ &= \left(\frac{\partial}{\partial x_1} + it_1\frac{\partial}{\partial x_2} + \cdots + it_{n-1}\frac{\partial}{\partial x_n}\right)^k \overline{(x_1 + is_1x_2 + \cdots + is_{n-1}x_n)^k}, \end{aligned}$$

which, by the previous lemma, equals

$$\begin{aligned} &k!\left[\left(\frac{\partial}{\partial x_1} + it_1\frac{\partial}{\partial x_2} + \cdots + it_{n-1}\frac{\partial}{\partial x_n}\right)\overline{(x_1 + is_1x_2 + \cdots + is_{n-1}x_n)}\right]^k \\ &= k!(1 + s_1t_1 + \cdots + s_{n-1}t_{n-1})^k, \end{aligned}$$

which, by the Funk–Hecke theorem [4, p. 247], equals

$$\sum_{j=1}^{d_n^k} \lambda_j P_j(s)P_j(t), \qquad \lambda_j \in \mathbb{R}.$$

On the other hand, by part (1) of Theorem 1,

$$\begin{aligned} (2)\quad &((x_1 + it_1x_2 + \cdots + it_{n-1}x_n)^k, (x_1 + is_1x_2 + \cdots + is_{n-1}x_n)^k) \\ &= \left(\sum_{j=1}^{d_n^k} Y_k^j(x)P_j(t), \sum_{l=1}^{d_n^k} Y_k^l(x)P_l(s)\right) \\ &= \sum_{j=1}^{d_n^k}\sum_{l=1}^{d_n^k}(Y_k^j, Y_k^l)P_j(t)P_l(s). \end{aligned}$$

Comparing (1) and (2) and using the linear independence of $\{P_j\}_{j=1}^{d_n^k}$, we see that $(Y_k^j, Y_k^l) = 0$ when $j \neq l$, completing the proof.

The generating function given by Theorem 1 can be used to construct the spherical harmonics in rectangular coordinates. Also, Fryant [2] has applied this theorem to give certain integral representations of harmonic functions in $\mathbb{R}^n$.

Acknowledgments. I thank Professor Fryant for introducing me to his generating function and the referee for simplifying the proof of Lemma 1.

REFERENCES

[1] A. FRYANT, *Inductively generating the spherical harmonics*, SIAM J. Math. Anal., 22 (1991), pp. 268–271.

[2] ———, *Integral operators for harmonic functions*, in The Mathematical Heritage of C. F. Gauss, G. Rassias, ed., World Scientific, Singapore, 1991.

[3] R. HOWE AND E.C.TAN, *Non-Abelian Harmonic Analysis*, Springer-Verlag, New York, Berlin, Heidelberg, 1992.

[4] A. ERDÉLYI, *Higher Transcendental Functions, Vol.* 2, McGraw–Hill, New York, 1953.

SIAM J. MATH. ANAL.
Vol. 26, No. 6, pp. 1647 1659, November 1995

015

NEW BOUNDS FOR HAHN AND KRAWTCHOUK POLYNOMIALS*

HOLGER DETTE†

Abstract. New identities for the sum of squares for the Hahn and Krawtchouk polynomials orthogonal on the set $\{0,\dots,N\}$ are derived which generalize the trigonometric identity for the Chebyshev polynomials of the first and second kind. These results are applied to obtain conditions (on the degree of the polynomials) such that the polynomials are bounded (on the interval $[0,N]$) by their values at the points 0 and N. As special cases we obtain a discrete analogue of the trigonometric identity and bounds for the discrete Chebyshev polynomials of the first and second kind.

Key words. Hahn polynomials, Hahn–Eberlein polynomials, Krawtchouk polynomials, dual Hahn polynomials, discrete trigonometric identity

AMS subject classification. 33C45

1. Introduction.

The Hahn polynomials may be defined in terms of a hypergeometric series

$$Q_n(x,\alpha,\beta,N) = {}_3F_2\left(\begin{matrix}-n, & n+\alpha+\beta+1, & -x;\\ \alpha+1, & -N & \end{matrix}\;1\right)$$
$$= \sum_{k=0}^{n} \frac{(-n)_k(n+\alpha+\beta+1)_k(-x)_k}{k!\,(\alpha+1)_k(-N)_k} \qquad (n=0,\dots,N),$$

where $\alpha,\beta > -1$, $(a)_0 = 1, (a)_k = a(a+1)\dots(a+k-1)$. These polynomials are limiting cases of some general systems of orthogonal polynomials (see Hahn (1949)) and satisfy, for $n,m = 0,\dots,N$, the orthogonality relation

$$\sum_{x=0}^{N} \rho(x,\alpha,\beta,N)Q_m(x,\alpha,\beta,N)Q_n(x,\alpha,\beta,N) = \frac{\delta_{nm}}{\pi_n(\alpha,\beta,N)}, \tag{1.1}$$

where

$$\rho(x,\alpha,\beta,N) = \frac{\binom{x+\alpha}{x}\binom{N-x+\beta}{N-x}}{\binom{N+\alpha+\beta+1}{N}} \tag{1.2}$$

and

$$\pi_n(\alpha,\beta,N) = \frac{(-1)^n(-N)_n(\alpha+1)_n(\alpha+\beta+1)_n}{n!\,(N+\alpha+\beta+2)_n(\beta+1)_n}\;\frac{2n+\alpha+\beta+1}{\alpha+\beta+1}. \tag{1.3}$$

For some properties and applications of the Hahn polynomials we refer the reader to the work of Karlin and McGregor (1961), (1962), Gasper (1974), (1975), and Wilson

*Received by the editors September 14, 1993; accepted for publication March 23, 1994. This research was supported in part by the Deutsche Forschungsgemeinschaft.

†Institüt für Mathematische Stochastik, Technische Universität Dresden, Mommsenstrasse 13, 01062 Dresden, Germany.

(1970). The polynomials $Q_n(x,\alpha,\beta,N)$ can be seen as the discrete analogue of the Jacobi polynomials and most of the "classical" orthogonal polynomials can be obtained as limits from the Hahn polynomials when the parameters tend to infinity (see Gasper (1975)).

As an example we consider the Krawtchouk polynomials, which can be defined as the limit ($q = 1-p, p \in (0,1)$)

$$\begin{aligned} k_n(x,p,N) &= \lim_{t\to\infty} Q_n(x,pt,qt,N) = {}_2F_1(-n,-x,-N;1/p) \\ &= \sum_{k=0}^{n} \frac{(-n)_k(-x)_k}{k!\,(-N)_k}\left(\frac{1}{p}\right)^k \end{aligned} \tag{1.4}$$

and are orthogonal with respect to the jump function

$$\binom{N}{x} p^x (1-p)^{N-x}, \qquad x = 0,\dots,N \tag{1.5}$$

(see Krawtchouk (1929)).

In this paper we will discuss some new properties of the orthogonal polynomials with respect to measures (1.2) and (1.5). After presenting some preliminary results in §2 we present new identities for squares of Krawtchouk and Hahn polynomials in §§3 and 4 which generalize the trigonometric identity for the Chebyshev polynomials of the first and second kind. We will apply these results to obtain conditions (on the degree of the polynomials) such that the polynomials $Q_n(x,\alpha,\beta,N)$ and $k_n(x,p,N)$ are bounded on the interval $[0,N]$ by their values at the points 0 and N. For the Hahn polynomials these bounds extend and improve results of Zaremba (1975), while for the Krawtchouk polynomials it is shown that

$$|k_n(x,p,N)| \le \max\{|k_n(0,p,N)|, |k_n(N,p,N)|\} = \max\left\{1, \left(\frac{q}{p}\right)^n\right\}, \qquad x \in [0,N],$$

whenever the degree of the polynomial satisfies $n \le \frac{N}{2} + 1$. Similar results are also given for the dual Hahn polynomials and the Hahn–Eberlein polynomials.

2. Preliminaries. In this section we will briefly discuss some general aspects of orthogonal polynomials which will be needed in the following sections. The notation used here is that of Karlin and Shapely (1953) and Karlin and Studden (1966). Let ξ denote a probability measure on the interval $[0,N]$ with moments

$$c_j = \int_0^N x^j d\xi(x) \qquad (j = 0,\dots,N)$$

and let $P_\ell(x), Q_\ell(x), R_\ell(x), S_\ell(x)$ denote the orthonormal polynomials with respect to the measures $d\xi(x), x(N-x)d\xi(x), xd\xi(x)$, and $(N-x)d\xi(x)$, respectively. The leading coefficients of these polynomials can be expressed by ratios of the determinants

$$\begin{cases} \underline{D}_{2\ell}(\xi) = |(c_{i+j})_{i,j=0}^{\ell}|, & \overline{D}_{2\ell} = |(Nc_{i+j-1} - c_{i+j})_{i,j=1}^{\ell}|, \\ \underline{D}_{2\ell+1}(\xi) = |(c_{i+j+1})_{i,j=0}^{\ell}|, & \overline{D}_{2\ell+1} = |(Nc_{i+j} - c_{i+j+1})_{i,j=0}^{\ell}| \end{cases} \tag{2.1}$$

(see, e.g., Karlin and Studden (1966) p. 109). For a point $(c_1, \ldots, c_\ell)$ in the interior of the moment space

$$\mathcal{M}_\ell = \{(c_1, \ldots, c_\ell) | c_j = \int_0^N x^j d\xi(x) \text{ for some probability measure on } [0, N] \quad (j = 1, \ldots, l)\},$$

let $(c_1, \ldots, c_{\ell-1}, c_\ell^-)$ and $(c_1, \ldots, c_{\ell-1}, c_\ell^+)$ denote the boundary points of $\mathcal{M}_\ell$ corresponding to the lower and upper principal representation associated with the point $(c_1, \ldots, c_{\ell-1}) \in \text{int}\,(\mathcal{M}_{\ell-1})$ (see Karlin and Studden (1966), p. 55). It is well known (see, e.g., Karlin and Shapely (1953), p. 59) that the quantities c_ℓ^+ and c_ℓ^- can be expressed in terms of the determinants (2.1), that is,

$$c_\ell^+ = c_\ell + \frac{\overline{D}_\ell(\xi)}{\overline{D}_{\ell-2}(\xi)}, \qquad c_\ell^- = c_\ell - \frac{\underline{D}_\ell(\xi)}{\underline{D}_{\ell-2}(\xi)}, \qquad \ell \geq 1, \tag{2.2}$$

where we define $\underline{D}_{-1}(\xi) = \underline{D}_0(\xi) = \overline{D}_{-1}(\xi) = \overline{D}_0(\xi) = 1$ (note that the ratios in (2.2) are well defined because $(c_1, \ldots, c_{\ell-1}) \in \text{int}\,(\mathcal{M}_{\ell-1})$).

Throughout this paper we will make use of the determinants defined in (2.1), where the moment of highest order is replaced by $c_{2\ell}^+$ $(c_{2\ell+1}^+)$ in the determinants $\underline{D}_{2\ell}(\xi)$ $(\underline{D}_{2\ell+1}(\xi))$ and by $c_{2\ell}^-$ $(c_{2\ell+1}^-)$ in the determinants $\overline{D}_{2\ell}(\xi)$ $(\overline{D}_{2\ell+1}(\xi))$. The corresponding modified determinants are denoted by $\underline{D}_{2\ell}^+(\xi), \underline{D}_{2\ell+1}^+(\xi), \overline{D}_{2\ell}^-(\xi)$ and $\overline{D}_{2\ell+1}^-(\xi)$, respectively. Using representation (2.2), it is then easy to see that

$$\begin{cases} \underline{D}_j^+(\xi) = \underline{D}_j(\xi) + \dfrac{\underline{D}_{j-2}(\xi)}{\overline{D}_{j-2}(\xi)} \overline{D}_j(\xi), & j = 2\ell,\ 2\ell+1, \\ \overline{D}_j^-(\xi) = \overline{D}_j(\xi) - \dfrac{\overline{D}_{j-2}(\xi)}{\underline{D}_{j-2}(\xi)} \underline{D}_j(\xi), & j = 2\ell,\ 2\ell+1. \end{cases} \tag{2.3}$$

In a recent paper Dette (1993) established new identities for the orthonormal polynomials $P_\ell(x), Q_\ell(x), R_\ell(x)$, and $S_\ell(x)$ with respect to the measures $d\xi(x)$, $x(N-x)d\xi(x)$, $xd\xi(x)$, and $(N-x)d\xi(x)$, respectively. For example, it is shown that for any arbitrary probability measure ξ on the interval $[0, N]$, the corresponding orthonormal polynomials satisfy the identity

$$\begin{aligned} &\sum_{\ell=1}^{n-1} \frac{\underline{D}_{2\ell-1}(\xi)}{\overline{D}_{2\ell-1}(\xi)} \left[\frac{\overline{D}_{2\ell-2}(\xi)}{\underline{D}_{2\ell-2}(\xi)} - \frac{\overline{D}_{2\ell}(\xi)}{\underline{D}_{2\ell}(\xi)} \right] P_\ell^2(x) + \frac{\underline{D}_{2n-1}(\xi)\, \overline{D}_{2n-2}(\xi)\, \underline{D}_{2n}(\xi)}{\overline{D}_{2n-1}(\xi)\, \underline{D}_{2n-2}(\xi)\, \underline{D}_{2n}^+(\xi)} P_n^2(x) \\ &+ (N-x) \sum_{\ell=0}^{n-1} \frac{\overline{D}_{2\ell}(\xi)}{\underline{D}_{2\ell}(\xi)} \left[\frac{\underline{D}_{2\ell-1}(\xi)}{\overline{D}_{2\ell-1}(\xi)} - \frac{\underline{D}_{2\ell+1}(\xi)}{\overline{D}_{2\ell+1}(\xi)} \right] S_\ell^2(x) \\ &= 1 - x(N-x) \frac{\underline{D}_{2n-1}(\xi) \overline{D}_{2n}(\xi)}{\overline{D}_{2n-1}(\xi) \underline{D}_{2n}^+(\xi)} Q_{n-1}^2(x) \end{aligned} \tag{2.4}$$

(note that the identities were originally stated on the interval $[-1, 1]$ but can easily be transferred to arbitrary intervals). If $N = 1$ and

$$d\xi(x) = \frac{dx}{\pi\sqrt{x(1-x)}}$$

is the arcsin distribution, then it is straightforward to show that $\overline{D}_{2\ell}(\xi) = \underline{D}_{2\ell}(\xi) = (\frac{1}{2})^{\ell(2\ell+1)}, \overline{D}_{2\ell+1}(\xi) = \underline{D}_{2\ell+1}(\xi) = (\frac{1}{2})^{(\ell+1)(2\ell+1)}$ (see, e.g., Karlin and Studden (1966), p. 123). The polynomials $P_\ell(x)$ and $Q_\ell(x)$ are proportional to the Chebyshev polynomials of the first and second kind (on the interval $[0,1]$) and the identity (2.4) reduces to the trigonometric identity. In this sense (2.4) can be seen as an extension of the trigonometric identity for arbitrary orthogonal polynomials on compact intervals. For the Jacobi polynomials identities of the form (2.4) have been established in Dette (1993). To derive similar results for the Hahn and Krawtchouk polynomials we need explicit expressions for the determinants of the moment matrices corresponding to the jump functions in (1.2) and (1.5), which will be derived in the following sections.

3. Identities and bounds for Hahn polynomials. It follows from (1.1) and (1.3) that the jump function in (1.2) defines a (discrete) probability measure ξ_ρ on the set $\{0,\dots,N\}$ and the orthonormal polynomials with respect to the measure $d\xi_\rho(x)$ are given by $\sqrt{\pi_n(\alpha,\beta,N)}Q_n(x,\alpha,\beta,N)$ $(n=0,\dots,N)$. Using the elementary properties of the gamma function and (1.1), we obtain

$$\begin{aligned}
&\sum_{x=0}^{N} Q_m(x-1,\alpha+1,\beta+1,N-2)Q_n(x-1,\alpha+1,\beta+1,N-2)x(N-x)\rho(x,\alpha,\beta,N)\\
&= \sum_{x=0}^{N-2} Q_m(x,\alpha+1,\beta+1,N-2)Q_n(x,\alpha+1,\beta+1,N-2)\\
&\qquad \times \rho(x,\alpha+1,\beta+1,N-2)\frac{N(N-1)(\alpha+1)(\beta+1)}{(\alpha+\beta+2)(\alpha+\beta+3)}\\
&= \frac{N(N-1)(\alpha+1)(\beta+1)}{(\alpha+\beta+2)(\alpha+\beta+3)}\cdot\frac{\delta_{m,n}}{\pi_n(\alpha+1,\beta+1,N-2)},
\end{aligned}$$

which shows that the polynomials

$$\sqrt{\frac{(\alpha+\beta+2)(\alpha+\beta+3)}{N(N-1)(\alpha+1)(\beta+1)}\pi_n(\alpha+1,\beta+1,N-2)}\; Q_n(x-1,\alpha+1,\beta+1,N-2) \tag{3.1}$$

$(n=0,\dots,N-2)$ are orthonormal with respect to the measure $x(N-x)d\xi_\rho(x)$. Similarly, it can be shown that the orthonormal polynomials with respect to the measures $xd\xi_\rho(x)$ and $(N-x)d\xi_\rho(x)$ are given by

$$\sqrt{\frac{(\alpha+\beta+2)}{(\alpha+1)N}\pi_n(\alpha+1,\beta,N-1)}\; Q_n(x-1,\alpha+1,\beta,N-1) \tag{3.2}$$

$(n=0,\dots,N-1)$ and

$$\sqrt{\frac{(\alpha+\beta+2)}{(\beta+1)N}\pi_n(\alpha,\beta+1,N-1)}\; Q_n(x,\alpha,\beta+1,N-1) \tag{3.3}$$

$(n=0,\dots,N-1)$, respectively.

THEOREM 3.1. *For $\ell=0,\dots,N$ define $h_\ell(x,\alpha,\beta,N) = ((\alpha+1)_\ell/(\beta+1)_\ell)Q_\ell(x,\alpha,\beta,N)$; then the Hahn polynomials satisfy the following identities:*

(a) *For* $n = 0, \dots, N-1$,

$$\sum_{\ell=1}^{n-1} \frac{2\ell+\alpha+\beta+1}{N}\{(\alpha+\beta+1)(2\ell-N)+2\ell^2\}$$
$$\times \left\{\frac{(\alpha+\beta+2)_{\ell-1}}{(N+\alpha+\beta+2)_\ell}\binom{N}{\ell}h_\ell(x,\alpha,\beta,N)\right\}^2$$
$$+\left\{\binom{N-1}{n-1}\frac{(\alpha+\beta+2)_{n-1}}{(N+\alpha+\beta+2)_{n-1}}h_n(x,\alpha,\beta,N)\right\}^2$$
$$+(\beta-\alpha)(1-\frac{x}{N})\sum_{\ell=0}^{n-1}\frac{2\ell+\alpha+\beta+2}{(\beta+1)^2}$$
$$\times\left\{\binom{N-1}{\ell}\frac{(\alpha+\beta+2)_\ell}{(\alpha+\beta+N+2)_\ell}h_\ell(x,\alpha,\beta+1,N-1)\right\}^2$$
$$=1-\frac{x}{N}(1-\frac{x}{N})$$
$$\times\left\{\frac{(\alpha+\beta+2)_n}{(\alpha+\beta+N+2)_{n-1}}\binom{N-2}{n-1}\frac{h_{n-1}(x-1,\alpha+1,\beta+1,N-2)}{\beta+1}\right\}^2.$$

(b) *For* $n = 0, \dots, N-1$,

$$\sum_{\ell=1}^{n}\frac{2\ell+\alpha+\beta+1}{N}\{(\alpha+\beta+1)(2\ell-N)+2\ell^2\}$$
$$\times\left\{\binom{N}{\ell}\frac{(\alpha+\beta+2)_{\ell-1}}{(\alpha+\beta+N+2)_\ell}Q_\ell(x,\alpha,\beta,N)\right\}^2$$
$$+\frac{x}{N}\left\{\binom{N-1}{n}\frac{(\alpha+\beta+2)_n}{(\alpha+\beta+N+2)_n}\frac{\alpha+n+1}{\alpha+1}Q_n(x-1,\alpha+1,\beta,N-1)\right\}^2$$
$$+(\alpha-\beta)\frac{x}{N}\sum_{\ell=0}^{n-1}(2\ell+\alpha+\beta+2)$$
$$\times\left\{\binom{N-1}{\ell}\frac{(\alpha+\beta+2)_\ell}{(\alpha+\beta+N+2)_\ell}\frac{Q_\ell(x-1,\alpha+1,\beta,N-1)}{\alpha+1}\right\}^2$$
$$=1-\left(1-\frac{x}{N}\right)\left\{\binom{N-1}{n}\frac{(\alpha+\beta+2)_n}{(\alpha+\beta+N+2)_n}Q_n(x,\alpha,\beta+1,N-1)\right\}^2.$$

(c) *For* $n = 0, \dots, N-2$,

$$x\left(1-\frac{x}{N}\right)\sum_{\ell=0}^{n-1}\{(\alpha+\beta+\ell+2)(N-2\ell-2)-(\ell+1)N\}(2\ell+\alpha+\beta+3)$$
$$\times\left\{\frac{h_\ell(x-1,\alpha+1,\beta+1,N-2)}{(\beta+1)(N-1)}\right\}^2$$
$$+\frac{x}{N}(1-\frac{x}{N})\left\{\frac{(\alpha+\beta+2+n)(N-n-1)}{(\beta+1)(N-1)}h_n(x-1,\alpha+1,\beta+1,N-2)\right\}^2$$
$$+(\beta-\alpha)\left(1-\frac{x}{N}\right)\sum_{\ell=0}^{n}(2\ell+\alpha+\beta+2)\left\{\frac{h_\ell(x,\alpha,\beta+1,N-1)}{\beta+1}\right\}^2$$
$$=1-\{h_{n+1}(x,\alpha,\beta,N)\}^2.$$

(d) *For* $n = 0, \dots, N-1$,

$$\begin{aligned}
&\sum_{\ell=1}^{n} \frac{2\ell+\alpha+\beta+1}{N}\{(\alpha+\beta+1)(2\ell-N)+2\ell^2\} \\
&\quad \times \left\{ \frac{(\alpha+\beta+2)_{\ell-1}}{(\alpha+\beta+N+2)_\ell} \binom{N}{\ell} h_\ell(x,\alpha,\beta,N) \right\}^2 \\
&\quad + (\beta-\alpha)\left(1-\frac{x}{N}\right) \sum_{\ell=0}^{n-1} (2\ell+\alpha+\beta+2) \\
&\quad \times \left\{ \binom{N-1}{\ell} \frac{(\alpha+\beta+2)_\ell}{(\alpha+\beta+N+2)_\ell} \frac{h_\ell(x,\alpha,\beta+1,N-1)}{\beta+1} \right\}^2 \\
&\quad + (1-\frac{x}{N}) \left\{ \binom{N-1}{n} \frac{\beta+n+1}{\beta+1} \frac{(\alpha+\beta+2)_n}{(\alpha+\beta+N+2)_n} h_n(x,\alpha,\beta+1,N-1) \right\}^2 \\
&= 1 - \frac{x}{N} \left\{ \binom{N-1}{n} \frac{(\alpha+\beta+2)_n}{(\alpha+\beta+N+2)_n} h_n(x-1,\alpha+1,\beta,N-1) \right\}^2 .
\end{aligned}$$

Proof. We will only give a proof of identity (a) using the general result in (2.4). All other cases are treated similarly, where identity (2.4) has to be replaced by the corresponding results in Dette (1993). Observing (2.4), (3.1), (3.2), and (3.3), we have to find the determinants $\underline{D}_{2\ell}(\xi_\rho), \overline{D}_{2\ell}(\xi_\rho), \underline{D}_{2\ell-1}(\xi_\rho), \overline{D}_{2\ell-1}(\xi_\rho)$, where ξ_ρ is the probability measure corresponding to the jump function (1.2). But these determinants can easily be calculated from the leading coefficients of the orthonormal polynomials with respect to the measures $d\xi_\rho(x), x d\xi_\rho(x), (N-x)d\xi_\rho(x), x(N-x)d\xi_\rho(x)$ (see, e.g., Karlin and Studden (1966), p. 110). For example, the orthonormal polynomial with respect to the measure $x(N-x)d\xi_\rho(x)$ is the Hahn polynomial given in (3.1) and the leading coefficient is obtained from the definition of the Hahn polynomials in terms of the hypergeometric series (see §1). Thus, for the leading coefficient of the polynomial in (3.1) we have

$$\begin{aligned}
&\sqrt{\frac{(\alpha+\beta+2)(\alpha+\beta+3)}{N(N-1)(\alpha+1)(\beta+1)}} \pi_n(\alpha+1,\beta+1,N-2) \cdot \frac{(n+\alpha+\beta+3)_n}{(\alpha+2)_n(-N+2)_n} \\
&\quad = (-1)^n \cdot \sqrt{\frac{\overline{D}_{2n}(\xi_\rho)}{\overline{D}_{2n+2}(\xi_\rho)}}
\end{aligned}$$

or, equivalently (using (1.3)),

$$\frac{\overline{D}_{2n+2}(\xi_\rho)}{\overline{D}_{2n}(\xi_\rho)} = \frac{n!\,(\alpha+1)_{n+1}(\beta+1)_{n+1}(N+\alpha+\beta+2)_n(N-n-1)_{n+2}}{(\alpha+\beta+2)_{n+1}(n+\alpha+\beta+3)_n(n+\alpha+\beta+3)_{n+1}} .$$

Similarly, for the ratio of $\underline{D}_{2n}(\xi_\rho)$ and $\underline{D}_{2n-2}(\xi_\rho)$ we obtain

$$\frac{\underline{D}_{2n}(\xi_\rho)}{\underline{D}_{2n-2}(\xi_\rho)} = \frac{n!\,(\alpha+1)_n(\beta+1)_n(\alpha+\beta+N+2)_n(N-n+1)_n}{(\alpha+\beta+n+1)_{n+1}(\alpha+\beta+n+1)_n(\alpha+\beta+2)_{n-1}}$$

and a straightforward computation yields

$$\frac{\overline{D}_{2n}(\xi_\rho)}{\underline{D}_{2n}(\xi_\rho)} = \frac{(N-n)(\alpha+\beta+n+1)}{n(N+\alpha+\beta+n+1)} \frac{\overline{D}_{2n-2}(\xi)}{\underline{D}_{2n-2}(\xi)} = \frac{(N-n)_n(\alpha+\beta+2)_n}{n!(N+\alpha+\beta+2)_n} . \tag{3.4}$$

In the same way we find

$$\frac{\overline{D}_{2n-1}(\xi_\rho)}{\underline{D}_{2n-1}(\xi_\rho)} = \frac{(\alpha+1)_n}{(\beta+1)_n}, \quad \frac{\underline{D}_{2n}(\xi_\rho)}{\underline{D}^+_{2n}(\xi_\rho)} = \frac{n}{N}\frac{\alpha+\beta+N+n+1}{\alpha+\beta+2n+1} \tag{3.5}$$

and

$$\frac{\overline{D}_{2n}(\xi_\rho)}{\underline{D}^+_{2n}(\xi_\rho)} = \frac{(\alpha+\beta+2)_n(N-n)_n}{(n-1)!\,(N+\alpha+\beta+2)_{n-1}(\alpha+\beta+2n+1)N}, \tag{3.6}$$

where we have used representation (2.3) and (3.4). The orthonormal polynomials with respect to the measures $(N-x)d\xi_\rho(x)$ and $x(N-x)d\xi_\rho(x)$ are given by (3.3) and (3.1), and assertion (a) of Theorem 3.1 now follows from (2.4), (3.4), (3.5), (3.6) and straightforward but tedious algebra. □

The Jacobi polynomials $P_\ell^{(\alpha,\beta)}(x)$, orthogonal with respect to the (continuous) measure $(1-x)^\alpha(1+x)^\beta dx$ and with leading coefficient $2^{-\ell}\binom{2\ell+\alpha+\beta}{\ell}$, can be obtained as limits from the Hahn polynomials

$$P_n^{(\alpha,\beta)}(x) = \lim_{N\to\infty}\binom{n+\alpha}{\alpha}Q_n\left(N\frac{1-x}{2},\alpha,\beta,N\right), \tag{3.7}$$

and replacing x by $-x$, it is straightforward to show that for the limit (3.7), Theorem 3.1 gives the corresponding formulas for the Jacobi polynomials in Dette (1993). For these polynomials it is well known that $|P_n^{(\alpha,\beta)}(x)|$ is bounded by $\max\{|P_n^{(\alpha,\beta)}(-1)|, |P_n^{(\alpha,\beta)}(1)|\}$ $(n \in \mathbb{N})$ if $\max\{\alpha,\beta\} > -\frac{1}{2}$. An upper, but not necessarily sharp, bound for arbitrary parameters is given by Erdélyi, Magnus, and Nevai (1992). For the Hahn polynomials the situation is more complicated. Zaremba (1975) showed that

$$|Q_n(x,\alpha,\beta,N)| \le 1 \tag{3.8}$$

for $x = 0,\dots,N$ provided that $\alpha \ge \beta > -1$, $n(n+1) \le N$, and

$$\alpha^2+\beta^2-\alpha\beta+\alpha+\beta \ge 0. \tag{3.9}$$

In the following theorem we will give an alternative bound for these polynomials, where the restriction on the degree of the polynomials satisfying (3.8) depends on the parameters of the weight function (1.2) and the inequality holds for all $x \in [0,N]$.

THEOREM 3.2. *Let* $\alpha+\beta > -1$ *and*

$$n(\alpha,\beta,N) := -\frac{1}{2}\{(\alpha+\beta-1) - \sqrt{(\alpha+\beta+1)(\alpha+\beta+2N+1)}\}; \tag{3.10}$$

then the nth Hahn polynomial satisfies the inequality

$$|Q_n(x,\alpha,\beta,N)| \le \max\left\{1, \frac{(\beta+1)_n}{(\alpha+1)_n}\right\} = \max\{|Q_n(0,\alpha,\beta,N)|,\ |Q_n(N,\alpha,\beta,N)|\}$$

for all $x \in [0,N]$ *and all* $n \le n(\alpha,\beta,N)$.

Proof. The second identity follows from Karlin and McGregor (1961) and equations (1.13) and (1.14). Let $\beta \ge \alpha$ and $\alpha+\beta > -1$, by (3.10) all terms on the left-hand

side of the identity in Theorem 3.1(c) are positive, which yields (here we replace n by $n-1$ in Theorem 3.1(c))

$$|Q_n(x,\alpha,\beta,N)| \leq \frac{(\beta+1)_n}{(\alpha+1)_n}$$

for all $x \in [0,N]$. If $\alpha \geq \beta$ we use the symmetry relation

$$Q_n(x,\alpha,\beta,N) = (-1)^n \frac{(\beta+1)_n}{(\alpha+1)_n} Q_n(N-x,\beta,\alpha,N)$$

(see, e.g., Nikifarov, Suslov, and Uvarov (1991), eq. (2.4.18), or Karlin and McGregor (1961), eq. (1.15), but note that both references use a different notation) and from the first part of the proof we obtain

$$|Q_n(x,\alpha,\beta,N)| = \left|\frac{(\beta+1)_n}{(\alpha+1)_n} Q_n(N-x,\beta,\alpha,N)\right| \leq 1$$

for all $x \in [0,N]$. This completes the proof of the theorem. □

Remark 3.3. Zaremba (1975) proved (3.8) for $\alpha \geq \beta > -1$ satisfying (3.9), $n(n+1) \leq N$, but only for the integers $x = 0,\ldots,N$, while Theorem 3.2 gives the sup-norm of the Hahn polynomials for all $\alpha+\beta > -1$. By restricting on the set $\{0,1,\ldots,N\}$ and $\alpha \geq \beta > -1$, Zaremba's bound on the degree of the polynomials (such that (3.8) is satisfied) is comparable with (3.10). If $\alpha = \beta = 0$, we obtain from Zaremba (1975) that (3.8) holds for all $n \leq (-1+\sqrt{4N+1})/2$, while Theorem 3.2 establishes the (for $N \geq 13$ weaker) bound $(1+\sqrt{2N+1})/2$. This can be explained by the fact that Zaremba's approach is directly related to the discrete Legendre polynomials $Q_n(x,0,0,N)$ (and to the integers $\{0,\ldots,N\}$) and the general case is obtained using a projection formula and results of Askey and Gasper (1971) (for this step the condition (3.9) is used). However, in most cases Theorem 3.2 will provide a better bound on the degree of the Hahn polynomials such that (3.8) is satisfied. Furthermore, condition (3.9) is not needed for establishing these bounds. For example, if $\alpha+\beta \geq 1$ and $N \geq 3$, then it is easy to see that $(-1+\sqrt{4N+1})/2 \leq n(\alpha,\beta,N)$ and, consequently, Theorem 3.2 gives a better bound on the degree of the polynomials, compared to the results of Zaremba (1975). Moreover, if $n \leq n(\alpha,\beta,N)$, (3.8) is satisfied for all $x \in [0,N]$. As a further example consider the case $\alpha = \beta > -\frac{1}{2}$ and $\beta(\beta+2) < 0$, then (3.9) is not satisfied and Zaremba's results cannot be applied. However, we readily obtain from Theorem 3.2 that (3.8) holds for all $x \in [0,N]$ whenever $n \leq \{-(2\beta-1)+\sqrt{(2\beta+1)(2\beta+1+2N)}\}/2$.

Zaremba (1975) also considered the example

$$Q_n\left(2,-\frac{1}{2},-\frac{1}{2},n^2\right) = -\frac{5}{3} \qquad (n \geq 2) \tag{3.11}$$

to show that condition (3.9) cannot be relaxed. In this case Theorem 3.2 is not applicable and (3.11) indicates that the Hahn polynomials $Q_n(x,\alpha,\beta,N)$ may not be bounded by their absolute values at the points 0 and N if $\alpha+\beta \leq -1$. Nevertheless, the following result provides a bound for these polynomials without a restriction on their degree.

THEOREM 3.4. *Let $\alpha+\beta \leq -1$ and $n \in \{0,\ldots,N-1\}$; then for all $x \in [0,N]$ the Hahn polynomials $Q_n(x,\alpha,\beta,N)$ satisfy the inequality*

(3.12)

$$|Q_n(x,\alpha,\beta,N)| \leq \max\left\{1, \frac{(\beta+1)_n}{(\alpha+1)_n}\right\} \cdot \frac{(\alpha+\beta+2+N)_{n-1}}{(\alpha+\beta+2)_{n-1}} \frac{(n-1)!}{(N-n+1)_{n-1}}.$$

Proof. Let $\beta \geq \alpha$; then by the assumptions all terms in the sums of Theorem 3.1(a) are positive. Consequently, we have

$$\left|\binom{N-1}{n-1} \frac{(\alpha+\beta+2)_{n-1}}{(\alpha+\beta+N+2)_{n-1}} h_n(x,\alpha,\beta,N)\right| \leq 1,$$

which is equivalent to (3.12) for $\beta \geq \alpha$. The case $\alpha \leq \beta$ is similar to the case in the proof of Theorem 3.2 and is therefore omitted. □

Remark 3.5. Note that in general the bound (3.12) cannot be improved. This follows readily from (3.11) for $N = 4, n = 2$ ($\alpha = \beta = -\frac{1}{2}$) because in this case the right-hand side of (3.12) is also given by $\frac{5}{3}$.

For $\alpha = \beta = -\frac{1}{2}$ we obtain the discrete analogue of the Chebyshev polynomials, which are of particular interest and considered in the following corollary. This result gives a "discrete" version of the trigonometric identity (part (a)).

COROLLARY 3.6. *Let* $T_n(x,N) = Q_n(x,-\frac{1}{2},-\frac{1}{2},N)$ *and* $U_n(x,N) = Q_n(x,\frac{1}{2},\frac{1}{2},N)$ *denote the discrete Chebyshev polynomials of the first and second kind, respectively; then we have the following for all* $x \in [0,N]$*:*

(a) *For* $n = 0,\ldots,N-1$,

$$-x\left(x-\frac{x}{N}\right)\sum_{\ell=0}^{n-1}(\ell+1)\left\{\frac{4(\ell+1)}{N-1}U_l(x-1,N-2)\right\}^2 + T_{n+1}^2(x,N) + \frac{x}{N}\left(1-\frac{x}{N}\right)\left\{\frac{2(n+1)(N-n-1)}{N-1}U_n(x-1,N-2)\right\}^2 = 1.$$

(b) *For* $n = 0,\ldots,N-1$,

$$|T_n(x,N)| \leq \prod_{j=1}^{n-1}\left(1+\frac{n}{N-n+j}\right).$$

(c) *For* $0 \leq n \leq \sqrt{N+1}$,

$$|U_n(x,N)| \leq 1.$$

Remark 3.7. Observing that the Jacobi polynomials can be obtained as the limit (3.7) from the Hahn polynomials and using formula (4.17) in Szegö (1975), it is easy to see that part (a) Corollary 3.6 yields ($N \to \infty$ $x = \frac{N}{2}(1-z)$) the trigonometric identity $(1-z^2)U_n^2(z)+T_{n+1}^2(z) = 1$ for the Chebyshev polynomial of the first and second kind while, parts (b) and (c) establish the bounds $|T_n(z)| \leq 1, |U_n(z)| \leq n+1$ ($z \in [-1,1]$) for these polynomials (note that $\lim_{N\to\infty} U_n(\frac{N}{2}(1-z),N) = U_n(z)/(n+1)$).

We will conclude this section with a brief discussion of related results for the Hahn–Eberlein and the dual Hahn polynomials. The Hahn–Eberlein polynomials are obtained from the Hahn polynomials $Q_n(x,\alpha,\beta,N)$ for $\alpha < -N, \beta < -N$ (see, e.g., Rahman (1978) or Eberlein (1964)). For this choice the mass function in (1.2) still

defines a probability measure on $\{0,\dots,N\}$ and, consequently, the orthogonal polynomials $Q_n(x,\alpha,\beta,N)$ with respect to this measure are well defined and called Hahn–Eberlein polynomials. These polynomials have some applications in coding theory (see, e.g., Sloane (1975)). Obviously, the identities of Theorem 3.1 can be extended to the region $\alpha < -N, \beta < -N$ and as a consequence we obtain the following bound for the Hahn–Eberlein polynomials.

THEOREM 3.8. *Let* $\alpha < -N, \beta < -N, \alpha+\beta < -2N-1$, *and*

$$\tilde{n}(\alpha,\beta,N) = -\frac{1}{2}\{(\alpha+\beta-1) + \sqrt{(\alpha+\beta+1)(\alpha+\beta+1+2N)}\}.$$

For all $x \in [0,N]$ *and* $n \leq \tilde{n}(\alpha,\beta,N)$ *the Hahn–Eberlein polynomials* $Q_n(x,\alpha,\beta,N)$ *satisfy the inequality*

$$|Q_n(x,\alpha,\beta,N)| \leq \max\left\{1, \frac{(\beta+1)_n}{(\alpha+1)_n}\right\} = \max\{|Q_n(0,\alpha,\beta,N)|, |Q_n(N,\alpha,\beta,N)|\}.$$

The dual Hahn polynomials $R_k(x,\alpha,\beta,N)$ $(\alpha,\beta > -1)$ are related to the Hahn polynomials by the equation

$$R_k(x(x+\alpha+\beta+1)) \;=\; Q_x(k,\alpha,\beta,N)$$

$(k,x = 0,\dots,N)$ and are orthogonal on the interval $[0, N(N+\alpha+\beta+1)]$. For a detailed discription of these polynomials including the recurrence relation and the orthogonality relation we refer the reader to the work of Karlin and McGregor (1961). By an analysis similar to the one in the proof of Theorem 3.2 we obtain the following bound for these polynomials.

THEOREM 3.9. *Let* $\alpha,\beta > -1$, $N+1+\beta-\alpha \geq 0$, *and*

$$n^*(\alpha,\beta,N) \;=\; \frac{1}{2}\min\{N+2,\; N+1+\beta-\alpha\}.$$

If $n \leq n^*(\alpha,\beta,N)$ *then for all* $x \in [0,N(N+\alpha+\beta+1)]$ *the dual Hahn polynomial* $R_n(x,\alpha,\beta,N)$ *satisfies the inequality*

$$|R_n(x,\alpha,\beta,N)| \leq \frac{(N+1+\beta-n)_n}{(\alpha+1)_n} \;=\; |R_n(N(N+\alpha+\beta+1),\alpha,\beta,N)|.$$

Proof. Let ξ_D denote the measure that puts masses

$$\xi_D(\lambda_x) = \pi_x(\alpha,\beta,N)\rho(0)$$

at the points $\lambda_x = x(x+\alpha+\beta+1)$ $(x = 0,\dots,N)$, where $\pi_x(\alpha,\beta,N)$ and $\rho(x) = \rho(x,\alpha,\beta,N)$ are defined in (1.3) and (1.2), respectively. By the results of Karlin and McGregor (1961) (equation (1.20)) it follows that ξ_D defines a probability measure on the interval $[0,N(N+\alpha+\beta+1)]$ and the orthonormal polynomials with respect to $d\xi_D(x)$ are given by

$$P_l(x) \;=\; \sqrt{\frac{\rho(l)}{\rho(0)}}R_l(x,\alpha,\beta,N) \qquad (l = 0,\dots,N). \tag{3.13}$$

According to Theorem 3.1 in Dette (1993), it follows that the orthonormal polynomials $P_l(x)$, $Q_l(x)$, and $S_l(x)$, with respect to the measures $d\xi_D(x)$, $x[N(N+\alpha+\beta+1)-x]d\xi_D(x)$, and $[N(N+\alpha+\beta+1)-x]d\xi_D(x)$, satisfy the identity
(3.14)

$$\begin{aligned} x[N(N+\alpha+\beta+1)-x]&\sum_{\ell=0}^{n-1}\frac{\underline{D}_{2\ell+1}(\xi_D)}{\overline{D}_{2\ell+1}(\xi_D)}\left[\frac{\underline{D}_{2\ell}(\xi_D)}{\overline{D}_{2\ell}(\xi_D)}-\frac{\underline{D}_{2\ell+2}(\xi_D)}{\overline{D}_{2\ell+2}(\xi_D)}\right]Q_\ell^2(x)\\ &+x[N(N+\alpha+\beta+1)-x]\frac{\underline{D}_{2n}(\xi_D)}{\overline{D}_{2n}(\xi_D)}\frac{\underline{D}_{2n+1}(\xi_D)}{\overline{D}_{2n+1}(\xi_D)}\frac{\overline{D}_{2n+2}(\xi_D)}{\overline{D}_{2n+2}(\xi_D)}Q_n^2(x)\\ &+[N(N+\alpha+\beta+1)-x]\sum_{\ell=0}^{n}\frac{\underline{D}_{2\ell}(\xi_D)}{\overline{D}_{2\ell}(\xi_D)}\left[\frac{\underline{D}_{2\ell-1}(\xi_D)}{\overline{D}_{2\ell-1}(\xi_D)}-\frac{\underline{D}_{2\ell+1}(\xi_D)}{\overline{D}_{2\ell+1}(\xi_D)}\right]S_\ell^2(x)\\ &=1-\frac{\underline{D}_{2n+1}(\xi_D)\underline{D}_{2n+2}(\xi_D)}{\overline{D}_{2n+1}(\xi_D)\overline{D}_{2n+2}(\xi_D)}P_{n+1}^2(x) \end{aligned}$$

$(n=0,\dots,N-1)$. Using reasoning similar to that in the proof of Theorem 3.1, for the ratios of the determinants in (3.14) we obtain

$$\frac{\underline{D}_{2l}(\xi_D)}{\overline{D}_{2l}(\xi_D)}-\frac{\underline{D}_{2l+2}(\xi_D)}{\overline{D}_{2l+2}(\xi_D)}=\frac{l!}{(N-l-1)_{l+1}}(N-2l-2)\quad(l=0,\dots,n-1),$$

$$\frac{\underline{D}_{2l-1}(\xi_D)}{\overline{D}_{2l-1}(\xi_D)}-\frac{\underline{D}_{2l+1}(\xi_D)}{\overline{D}_{2l+1}(\xi_D)}=\frac{(\alpha+1)_l}{(N+\beta-l)_{l+1}}(N-1+\beta-\alpha-2l)\quad(l=0,\dots,n),$$

and

$$\begin{aligned} &\frac{\underline{D}_{2n+1}(\xi_D)\underline{D}_{2n+2}(\xi_D)}{\overline{D}_{2n+1}(\xi_D)\overline{D}_{2n+2}(\xi_D)}P_{n+1}^2(x)\\ &\quad=\frac{(\alpha+1)_{n+1}(n+1)!}{(N+\beta-n)_{n+1}(N-n)_{n+1}}\frac{\rho(n+1)}{\rho(0)}R_{n+1}^2(x,\alpha,\beta,N)\\ &\quad=\left(\frac{(\alpha+1)_{n+1}}{(N+\beta-n)_{n+1}}R_{n+1}(x,\alpha,\beta,N)\right)^2, \end{aligned}$$

where we have used (3.13) and (1.2) in the last identity. By the assumptions of the theorem all terms on the left-hand side in (3.14) are positive and the assertion follows from

$$R_n(N(N+\alpha+\beta+1))=Q_N(n,\alpha,\beta,N)\;=\;(-1)^n\frac{(N+1+\beta-n)_n}{(\alpha+1)_n},$$

which can easily be proved by an induction argument. □

4. Krawtchouk polynomials. In this section we will apply the results of §§2 and 3 to obtain similar results for the Krawtchouk polynomials. We will mainly use the representation (1.4) of $k_n(x,p,N)$ as the limit of the Hahn polynomials $Q_n(x,\alpha,\beta,N)$ when $\alpha=pt,\beta=qt$, and $t\to\infty$. By this relation the following results are immediate consequences of Theorems 3.1 and 3.2.

THEOREM 4.1. *For $\ell=0,\dots,N$ define $\tilde{k}_\ell(x,p,N)=\binom{N}{\ell}(\frac{p}{q})^\ell k_\ell(x,p,N)$. The Krawtchouk polynomials satisfy the following identities:*

(a) *For* $n = 0, \dots, N-1$,

$$\sum_{\ell=1}^{n-1} \left(\frac{2\ell}{N} - 1\right) \{\tilde{k}_\ell(x,p,N)\}^2 + \left\{\frac{n}{N}\tilde{k}_n(x,p,N)\right\}^2 + \left(1 - \frac{x}{N}\right)\frac{q-p}{q^2}\sum_{\ell=0}^{n-1}\{\tilde{k}_\ell(x,p,N-1)\}^2 = 1 - \frac{x}{N}\left(1-\frac{x}{N}\right)\left\{\frac{k_{n-1}(x-1,p,N-2)}{q}\right\}^2.$$

(b) *For* $n = 0, \dots, N-1$,

$$\sum_{\ell=1}^{n}\left(\frac{2\ell}{N}-1\right)\left\{\binom{N}{\ell}k_\ell(x,p,N)\right\}^2 + \frac{x}{N}\left\{\binom{N-1}{n}k_n(x-1,p,N-1)\right\}^2 + \frac{p-q}{p^2}\frac{x}{N}\sum_{\ell=0}^{n-1}\left\{\binom{N-1}{\ell}k_\ell^2(x-1,p,N-1)\right\}^2 = 1 - \left(1-\frac{x}{N}\right)\left\{\binom{N-1}{n}k_n(x,p,N-1)\right\}^2.$$

(c) *For* $n = 0, \dots, N-2$,

$$x\left(1-\frac{x}{N}\right)\sum_{\ell=0}^{n-1}(N-2\ell-2)\left\{\frac{p^\ell}{q^{\ell+1}}\frac{k_\ell(x-1,p,N-2)}{(N-1)}\right\}^2 + \frac{q-p}{q^2}\left(1-\frac{x}{N}\right)\sum_{\ell=0}^{n}\left\{\frac{p^\ell}{q^\ell}k_\ell(x,p,N-1)\right\}^2 + \frac{x}{N}\left(1-\frac{x}{N}\right)\left\{\frac{(N-n-1)p^n k_n(x-1,p,N-2)}{q^{n+1}(N-1)}\right\}^2 = 1 - \left\{\frac{p^{n+1}}{q^{n+1}}k_{n+1}(x,p,N)\right\}^2.$$

(d) *For* $n = 0, \dots, N-1$,

$$\sum_{\ell=1}^{n}\left(\frac{2\ell}{N}-1\right)\{\tilde{k}_\ell(x,p,N)\}^2 + \frac{q-p}{q^2}\left(1-\frac{x}{N}\right)\sum_{\ell=0}^{n-1}\{\tilde{k}_\ell(x,p,N-1)\}^2 + \left(1-\frac{x}{N}\right)\{\tilde{k}_n(x,p,N-1)\}^2 = 1 - \frac{x}{N}\{\tilde{k}_n(x-1,p,N-1)\}^2.$$

THEOREM 4.2. *Let* $n \le \frac{N}{2} + 1$, *then the nth Krawtchouk polynomial* $k_n(x,p,N)$ *for all* $x \in [0,N]$ *satisfies the inequality*

$$|k_n(x,p,N)| \le \max\left\{1, \left(\frac{q}{p}\right)^n\right\} = \max\{|k_n(0,p,N)|, |k_n(N,p,N)|\}.$$

Remark 4.3. It should be noted that the bound $\frac{N}{2} + 1$ for the degree of the Krawtchouk polynomials in the preceeding theorem could be improved by using a

positive linearization theorem for the Krawtchouk polynomials [see, e.g., Dunkl and Ramirez (1974)].

Acknowledgments. Parts of this paper were written while the author was visiting the University of Göttingen and Purdue University, West Lafayette. The author thanks the Institut für Mathematische Stochastik and the Department of Statistics for their hospitality and the Deutsche Forschungsgemeinschaft for the financial support that made a visit to the United States possible. I am also indebted to Dick Askey and George Gasper for their helpful comments and help with the references. It was George Gasper who convinced me that identities of the form (2.4) may give useful bounds for orthogonal polynomials, while Dick Askey informed me of the results of S. K. Zaremba.

REFERENCES

R. Askey and G. Gasper (1971), *Jacobi polynomial expansions of Jacobi polynomials with non-negative coefficients*, Proc. Cambridge Philos. Soc., 70, pp. 245–255.

H. Dette (1993), *New identities for orthogonal polynomials on compact intervals*, J. Math. Anal. Appl., 179, pp. 547–573.

C. Dunkl and D. E. Ramirez (1974), *Krawtchouk polynomials and the symmetrization of hypergroups*, SIAM J. Math. Anal., 5, pp, 351–366.

P. J. Eberlein (1964), *A two parametric test matrix*, Math. Comput., 18, pp. 296–298.

T. Erdélyi, A. P. Magnus, and P. Nevai (1992), *Generalized Jacobi weights, Christoffel functions, and Jacobi Polynomials*, Ohio State Mathematical Research Institute preprint, 92-29.

G. Gasper (1974), *Projection formulas for orthogonal polynomials of a discrete variable*, J. Math. Anal. Appl., 45, pp. 176–198.

——— (1975), *Positivity and special functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, pp. 375–433.

W. Hahn, *Über Orthogonalpolynome, die q-Differenzengleichungen genügen*, Math. Nachr., 2, pp. 4–34.

S. Karlin and J. McGregor (1961), *The Hahn polynomials, formulas and an application*, Scripta Math. 26, pp. 33–46.

——— (1962), *On a genetics model of Moran*, Proc. Cambridge Philos. Soc., 58, pp. 299–311.

S. Karlin and L. S. Shapeley (1953), *Geometry of Moment Spaces*, Amer. Math. Soc. Memoir No 12, American Mathematical Society, Providence, RI.

S. Karlin and W. J. Studden (1966), *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience, New York.

M. Krawtchouk (1929), *Sur une généralisation des polynomés d' Hermite*, Comptes Rendus de l' Académie des Sciences, Paris, 189, pp. 620–622.

A. F. Nikiforov, S. K. Suslov, and V. B. Uvarov (1991), *Classical Orthogonal Polynomials of a Discrete Variable*, Springer-Verlag, New York.

M. Rahman (1978), *A positive kernel for Hahn–Eberlein polynomials*, SIAM J. Math. Anal., 9, pp. 891–905.

N. J. A. Sloane (1975), *An introduction to association schemes and coding theory*, in Theory and Application of Special Functions R. Askey, ed., Academic Press, New York, pp. 225-260.

G. Szegö (1975), *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ., Vol. 23, American Mathematical Society, New York.

M. W. Wilson (1970), *On the Hahn polynomials*, SIAM J. Math. Anal., 1, pp. 131–139.

S. K. Zaremba (1975), *Some properties of polynomials orthogonal over the set* $\langle 1, 2, \ldots, N \rangle$, Ann. Mat. Pura Appl., 105, pp. 333–345.